



저작자표시-비영리-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

약학박사학위논문

약물 리포지셔닝과 약물 표적 동정을 위한
약물학에서의 네트워크 분석

**Network Analysis in Pharmacology for Drug
Repositioning and Drug Target Identification**

2014 년 2 월

서울대학교 대학원

약학과 약학전공

배 태 정

약물 리포지셔닝과 약물 표적 동정을 위한
약물학에서의 네트워크 분석

**Network Analysis in Pharmacology for Drug
Repositioning and Drug Target Identification**

지도교수 김 성 훈

이 논문을 약학박사학위논문으로 제출함
2013 년 10 월

서울대학교 대학원
약학과 약학전공
배 태 정

배태정의 약학박사학위논문을 인준함
2013 년 12 월

위	원	장	<u>신 영 기</u>	(인)
부	위	원	<u>장 한 병 우</u>	(인)
위		원	<u>이 상 혁</u>	(인)
위		원	<u>황 대 희</u>	(인)
위		원	<u>김 성 훈</u>	(인)

ABSTRACT

Network biology represents and analyzes biological systems as complex networks consisting of various kinds of cellular molecules and relationships between them to understand cellular functions at system level. Viewing biological systems via the network concept may also help us improve drug discovery by revealing more complex aspect of drug action in cellular networks. To apply network analysis into drug discovery process, first of all, we gathered comprehensive knowledge from diverse disciplines related to drug discovery and integrated it into a pharmacological tripartite network database (PharmDB) consisting of drugs, targets, diseases and their relationships. Secondly, we developed “shared neighborhood scoring algorithm”, a new method proper to analyze our pharmacological tripartite network since existing network analysis approaches use various projection methods to convert a multi-partite network into several mono-partite networks causing information loss. By combining PharmDB with the shared neighborhood scoring algorithm, we can explore the pharmacological tripartite network to identify new drug targets, to design drug repositioning and new drug combination, and to predict potential side effect of a drug. Thirdly, in addition to network topology analysis based on prior knowledge, we tried network inference, a kind of data-driven approach to utilize large-scale gene expression profiles into network analysis. We constructed a transcriptional network using ARACNE as a network inference algorithm and analyzed it to extract master regulators or transcription factors (TFs) for known prognostic signature genes of colorectal cancer. All of this works

represents that network biology can be a fully integrated solution to better, more efficient drug discovery and development.

Keywords

Pharmacological database, data integration, tripartite network, drug target, network topology, shared neighborhood scoring algorithm, drug repositioning, gene signature, colorectal cancer, transcriptional network, network inference

Student number: 2009-30464

CONTENTS

Abstract	1
Contents	3
List of figures and tables	5
List of abbreviations	7
Introduction	9

Chapter I

Development of a Novel Algorithm to Search an Integrated Pharmacological Network for Drug Repositioning	14
--	-----------

Abstract	15
Introduction	16
Results and discussions	18
Conclusions	30
Methods	31
Abbreviations	38
References	39
Figures and Tables	45

Chapter II

Identification of Upstream Regulators for Prognostic Expression Signature Genes in Colorectal Cancer	67
---	-----------

Abstract	68
Introduction	70

Results and discussions -----	73
Conclusions -----	81
Methods -----	82
Abbreviations -----	84
References -----	85
Figures and Tables -----	91
국문초록 -----	107

LIST OF FIGURES AND TABLES

Chapter I

Development of a Novel Algorithm to Search an Integrated Pharmacological Network for Drug Repositioning

Figure I-1. Data model and architecture of PharmDB -----	45
Figure I-2. Data integration process of PharmDB -----	47
Figure I-3. Different output patterns in PharmDB -----	49
Figure I-4. Data statistics of PharmDB -----	51
Figure I-5. Possible applications of PharmDB -----	53
Figure I-6. Missing links in the PharmDB network -----	54
Figure I-7. Shared neighborhood node distribution -----	55
Figure I-8. Non-linear regression results for extracting connecting probability function -----	56
Figure I-9. Connecting probability model for SN score -----	57
Figure I-10. ROC curves of the shared neighborhood scoring algorithm ---	59
Figure I-11. AUC changes by random sampling -----	60
Figure I-12. AUC changes by random deletion -----	61
Figure I-13. A workflow for drug repositioning -----	62
Figure I-14. Experimental validation of the effect of TBZT against SCC --	63
Table I-1. Summary of PharmDB data sources -----	65
Table I-2. List of drug candidates and drug-associated proteins -----	66

Chapter II

Identification of Upstream Regulators for Prognostic Expression Signature Genes in Colorectal Cancer

Figure II-1. Overall pipeline of upstream regulator inference -----	91
Figure II-2. The transcriptional network between the top 10 TFs and the signature genes -----	93
Figure II-3. Correlation between the upstream TFs and their target genes -----	96
Figure II-4. Expression patterns of the selected marker genes between the good and the poor prognostic group -----	97
Figure II-5. The prediction performance of the selected prognostic markers --	98
Table II-1. 85 Signature genes -----	99
Table II-2. Master Regulator Candidates -----	100
Table II-3. Prognostic Effect Summary -----	102
Table II-4. Overall statistics of 13 TFs, union of TF_{MRA} and $TF_{MRA+SLR}$ ---	104
Table II-5. Top 10 TF-Target list of MRA and MRA+SLR method -----	105

LIST OF ABBREVIATIONS

TTD: Therapeutic Target Database

MeSH: Medical Subject Headings

ICD: International Classification of Diseases

PTN: pharmacological tripartite network

KEGG: Kyoto Encyclopedia of Genes and Genomes

PPI: protein-protein interaction

BIND: Biomolecular INteraction Network Database

HPRD: Human Protein Reference Database

DIP: the Database of Interacting Proteins

MINT: the Molecular INteraction database

DOUT: drug of unknown target

SNS: shared neighborhood scoring

SN score: shared neighborhood score

PharmGKB: The Pharmacogenomics Knowledge Base

OMIM: Online Mendelian Inheritance in Man

GAD: Genetic Association Database

ROC: receiver operating characteristic

AUC: area under curve

AZA: Acetazolamide

CA: carbonic anhydrase

SCC: squamous cell carcinoma

TBZT: thia-benzthiazide

KS test: Kolmogorov-Smirnov test

MRA: master regulator analysis

SLR: stepwise linear regression

AML: acute myeloid leukemia

GBM: glioblastoma

TCF: T-cell factor

LEF: lymphoid enhancer factor

INTRODUCTION

Cells are operated via a network of diverse cellular molecules including proteins, metabolic and signaling pathways, and transcriptional or translational regulation, in which the complexity outgrows the simple sum of all the components. To understand working principle of living organisms at system level, it is of a great importance to build a holistic model of the system taking into account the relationships of all the components with the associated functions. The system-level understanding of living organisms may help to develop predictable models of human diseases which can be used in drug discovery process. Despite the efforts to catalogue the whole network of cellular interacting molecules during last decades, it is still far from satisfaction mainly due to the complexity of the living organism and technical limitations to analyze the behavior of the components. However, thanks to the recent advance of high-throughput technology, various types of large-scale data become available to accumulate pathway information much faster than before.

To apply the information on biological pathways and protein interactions for drug discovery, pharmacological information should be also integrated. To integrate widely scattered pharmacological knowledge and databases and link those to biological network, we constructed a new database called PharmDB. This database is a comprehensive catalogue of integrated pharmacological network which consists of inter-relationships of drugs, targets and diseases and displays its own content in dynamic network view. In addition, the pharmacological network of PharmDB is further linked to

protein-protein interactions and biological signal pathways via therapeutic target proteins to expand the current knowledge on drug action and applied diseases. Although there are databases providing information on drugs and their relations to targets and diseases, there is no database in which all of three entities are integrated in a unified platform and their vertical and horizontal linkages are displayed in graphical view. This type of data presentation would help to get over the limitation of text-based knowledge and to overview the potential connection between the entities that seemingly look remote.

The pharmacological network of PharmDB is a tripartite network which contains three kinds of nodes, i.e., drug, target and disease. It is unquestionable that any integrated analysis of the tripartite network of PharmDB gives us the most abundant information to find important functional relations among the three kinds of nodes. Since multi-partite network is more difficult to be analyzed than mono-partite network, existing network analysis approaches include various projection processes to convert a multi-partite network to several mono-partite networks. However, any projection process of networks makes lose some important information especially on low-degree nodes. Therefore, development of novel methods is necessary to analyze intact multi-partite network without information loss by projection.

For this needs, we designed a new algorithm i.e. shared neighborhood scoring algorithm. The shared neighborhood score is a measure of the supporting evidence to connect two nodes with shared neighbors. In social networks, for example, two persons who have shared friends are much more likely to be also friends, as compared with two randomly chosen persons. Here, as we guess, the probability that the two persons are friends each other

is proportional to the number of their shared friends: the probability becomes high when their shared friends are many. Similarly the shared neighborhood score is also proportional to the number of shared neighbors of one pair of nodes. If the number is large, then the pair of nodes has a high possibility to be really connected each other. The shared neighborhood scoring algorithms are based on measuring this possibility and used to find any highly plausible link. This idea is thus similar to the approach of Swanson's ABC model

The shared neighborhood scoring algorithm is applicable to predict yet-unreported possible links between any two kinds of nodes in the pharmacological network of PharmDB to find novel drug targets, drug repositioning candidates, or novel target phenotypes. In case of finding novel drug targets, when one drug is connected with several diseases and many of them have relations with one specific protein in common, one can suggest that the protein may be a target of the drug. The probability that the protein becomes a target of the drug is proportional to the number of diseases which are commonly linked to both the drug and the protein. As there are more common diseases, the protein is more likely to be a novel target of the drug. We can thus find a highly plausible link between one drug and one protein through many commonly connected diseases. In similar manners, the shared neighborhood scoring algorithm can help us to gather new integrated information on finding new candidates such as new disease–target associations, new drug side-effects, or new drug efficacy and repositioning. We validated the utility of this algorithm to drug repositioning, by identifying a potential application of a hypertension drug, benzthiazide (TBZT), to induce lung cancer cell death.

Although PharmDB combined with the shared neighborhood scoring algorithm is a useful knowledge platform to find hidden links in pharmacological network, knowledge-based approaches have limitations in regard that they search only already-known information. In addition, the shared neighborhood scoring algorithm uses only topological structure of network to estimate plausibility of suggesting hidden links. This excludes conditional states or dynamic changes of cellular network. To complement these limitations of network topology analysis, we considered network inference using large-scale gene expression profiles. Network inference is a kind of data-driven approach to construct network structure between genes without prior knowledge. Among the various network inference algorithms, we adopted the ARACNE algorithm to construct a transcriptional network. Then we analyzed the transcriptional network to extract upstream regulators or transcription factors (TFs) for known prognostic signature genes of colorectal cancer.

To this end, a global regulatory network is modeled using ARACNE with 177 expression profiles from colorectal cancer. Then, we applied essentially the same method as the previous work by Carro and colleagues, which was originally applied to brain cancer (high grade glioblastoma). Through this process, we selected the top 10 TFs with the highest coverage of the prognostic signature genes and tested the performance for prognosis using an independent dataset. The selected top 10 TFs show a slightly better performance than the original prognostic signature, although the signature size dramatically reduced from 85 to 10, which is highly desirable for practical application. This result establishes the utility of upstream regulators as

potentially better prognostic marker than conventional expression signatures. The selected top 10 TFs included many known regulators for tumorigenic processes and the remaining TFs are also likely to provide clues to novel functions and the underlying regulatory mechanisms in colorectal cancer. This approach may be applicable to other types of cancer signatures such as metastasis and drug response, which, however, has not been thoroughly investigated by experiments.

Chapter I

Development of a Novel Algorithm to Search an Integrated Pharmacological Network for Drug Repositioning

ABSTRACT

There is a need to find alternative ways to address the issue of low productivity in drug discovery. In this study, we established a database we call “PharmDB” showing the integrated network of diseases, drugs, and associated proteins. We also developed an algorithm called “Shared Neighborhood Scoring (SNS) algorithm” to discover potential linkages in the integrated network and designed a logical workflow that suggests potential efficacy of the known drugs to the disease of interest. Using this SNS algorithm and the knowledge-based tool kit based on it, we identified the potential application of a hypertension drug, benzthiazide (TBZT), to induce cancer cell death.

Keywords

Pharmacological database, tripartite network, data integration, drug target, network topology, shared neighborhood scoring algorithm, drug repositioning

INTRODUCTION

Despite the technological advances in drug discovery, the approval of new drugs has remained stagnant in the past several years, resulting in the decline of productivity in the pharmaceutical industry. In this regard, the repositioning of currently available drugs to other indications is considered an alternative to save on drug development time and to minimize risks of failure(1). However, most of the previously repositioned drugs have resulted from the serendipitous observation of the unexpected efficacy and side effects of developing or on-market drugs. The goal of this study is to establish a knowledge platform that can logically suggest the repositioning of a known drug to a new indication. We designed an information tool that could rapidly generate suggestions about potential links among diseases, drugs, and proteins by compiling data that contain information on the binary linkages among these three components. The resulting integrated database was designated as “PharmDB” (Figure I-1A), a tripartite pharmacological network database consisting of three kinds of nodes, namely, human diseases, drugs, and proteins. The proteins in PharmDB include therapeutic targets, disease-associated proteins, and drug-metabolizing proteins.

A number of network analysis measures are proposed to draw useful insight from topological or structural properties of various complex networks. Most of these measures are applicable only to a monopartite network that consists only of one type of node. Therefore, a bipartite network with two kinds of nodes cannot be analyzed using these measures. To solve this problem, researchers have used projection methods that convert bipartite

networks into monopartite ones. Unfortunately, any projection method can result in information loss, especially in low-degree nodes(2,3). These projection methods distort many well-known network measures, such as average path length $\langle l \rangle$, average clustering coefficient $\langle C \rangle$, degree-dependent clustering coefficient $C(k)$, degree distribution $P(k)$, assortativity coefficient r , and degree-degree correlation coefficient $k_{nn}(k)$.

PharmDB is constructed as a tripartite network with three kinds of nodes, and thus the projection of the PharmDB tripartite network can become complex and lose information. To overcome these limits of the projection technique, we introduced a new method for analyzing bi- or tripartite networks without the use of any projection. We designed a new analyzing method called “shared neighborhood scoring algorithm” based on the missing links in the PharmDB tripartite network.

RESULTS AND DISCUSSIONS

PharmDB data structure and standard identifiers

Although there are active movements toward unifying and centralizing data using standard identifiers (4,5), numerous independent databases are getting published with their own identification system with few cross references to other collective databases such as Entrez Gene or PubChem (6,15). Although such independent databases could be useful and convenient in the specific domain of knowledge, it is difficult for general users to integrate them for more comprehensive investigation or extrapolation. Since drug discovery involves comprehensive knowledge of biology, chemistry, pharmacology and medicine, generation of an integrated and expandable database covering the data of those areas is in demand (4,8). For this reason, we decided to take standard identifiers that satisfy the following criteria. First, they should be used and supported by large scientific community. Second, they should have extensive cross-references and synonym tables which may link data to other databases. Third, the coverage of the database using the identifiers should be large. Fourth, the organization adopting them should have an active curation system that oversees the assignment of the unique identifier. These conditions would be necessary to ensure the stability of the assigned identifier without frequent changes.

With these conditions in mind, we chose to use GeneID (9) from Entrez Gene in NCBI as the identifier for proteins. GeneID is not only widely used, but also has a wide coverage of human genome and is more stable than accession ID. For drug identifier, we looked into the two databases: CAS

registry and PubChem (10). Although CAS registry, the identifier of which is CAS registry number, has the largest coverage of small molecules, it is privately owned, expensive and often available only through restricted interfaces (11). Thus, we decided to take the identification system of PubChem, which uses two kinds of identifiers, PubChem SID and PubChem CID. The former is initially assigned to all of the data submitted by a number of organizations such as KEGG and ChemDB. Therefore, the entries assigned with PubChem SID might have many redundancies and even erroneous entries. On the other hand, PubChem CID is assigned to unique structures which are reviewed and curated after the entries with PubChem SID are pre-clustered by structural similarity. For this reason, the entities with PubChem CID are more reliable and cleaner than those with only PubChem SID. Therefore, PharmDB uses mainly PubChem CID as drug identifier. For some entries (99 out of 2,697) that have only PubChem SIDs with no assigned PubChem CIDs, we temporarily adopted PubChem SID as a complementary identification system, which will be changed to PubChem CIDs as soon as they have the assignment. For disease identifier, we looked into MeSH (Medical Subject Headings) (12) of the National Library of Medicine (NLM) and ICD (International Classification of Diseases) published by World Health Organization (WHO) (13). Although ICD-10, the latest edition of ICD classification system, is used for all general epidemiological and many health management purposes, it does not provide any disease descriptions and a disease name synonym table. Since MeSH provides disease descriptions and synonyms on each descriptor, PharmDB decided to adopt MeSH Unique ID as the disease identifier.

Data integration process for PharmDB

The data for PharmDB were collected by integrating publicly available databases, such as TTD (14), DrugBank (15) and ChemBank (16). More specifically, drug-target relations were extracted from TTD and DrugBank while the relations of target-disease and disease-drug relations were obtained from TTD, and DrugBank and ChemBank, respectively. With these collected data, we created a conceptual graphical network that contains drug, target and disease as nodes and their inter-relations as edges. We defined this conceptual graphical network as the pharmacological tripartite network (PTN) (Figure I-1B). Once we obtained the raw data, we put them through a series of data clean up pipeline scripts to remove duplicates and assign standard identifiers to each entry (Figure I-2). During the clean-up process, pieces of annotation information for each entry were provided by the source databases, such as synonyms, descriptions and commonly used identifiers from a variety of biomedical databases were used to match protein entities with GeneIDs, drug entities with PubChem CIDs or PubChem SIDs. We also manually processed raw data obtained from TTD since it only provides synonyms and gene names for target proteins in contrast to DrugBank and ChemBank that also give commonly used identifiers of well recognized organizations such as NCBI or EBI. TTD also contain target proteins of pathogens and anti-infectives. Since PharmDB presently focuses on human diseases resulting from the disturbance of endogenous targets, we excluded the exogenous targets originated from various human pathogens. 811 disease entries obtained from TTD were also manually curated to match them with MeSH descriptor.

Cellular proteins such as receptors, enzymes and signal mediators constitute a majority of therapeutic targets and they are often linked to signaling and metabolic pathways via complex network with surrounding proteins. Thus, it would be informative to look into the proteins surrounding each individual target protein in order to get the insight into possible activities and side effect of a drug working to the target protein. For this, we first gathered the necessary data from the publicly available protein-protein interaction and signaling pathway databases as many as possible. KEGG pathway (17) contains detailed signaling network data. We combined comprehensive protein-protein interaction (PPI) data from Entrez Gene interaction database (9), which contains a collection of the HPRD, BIND, DIP databases (18) to KEGG signal pathways. They catalog experimentally determined PPIs which were automatically harvested using computational approaches and manually curated. We also included the experimental PPI data from the works of Lim (19), Rual (20), and Stelzl (21). In addition, we integrated the two predicted PPI data of human proteins: HomoMINT (22) and Rhodes *et al.* (23). HomoMINT was constructed through orthology mapping of protein interactions based on the MINT database, a large PPI database of model organisms. The data from Rhodes *et al.* were predicted through a probabilistic analysis based on the data of model organism interactome, protein domains, genome-wide gene expression and functional annotation. Although the predicted datasets might be less reliable than the experiment-based data, we thought they would be useful to temporarily fill up the missing links in human PPI network until more experimental data become available (24).

phExplorer: a network browser for PharmDB

Although data presentation in a textual tabular format can be convenient to give detailed information on each entity separately, it is not ideal to simultaneously describe the relations between proteins, pathways, diseases and drugs that are associated with the protein of interest (25,26). In this case, displaying the data in a graphical network format would be more helpful to envisage inter-connections between the extracted data (4). For this purpose, we developed a supplementary Java applet tool called phExplorer beta to graphically browse and investigate PharmDB data. In phExplorer, drug, protein and disease entities are represented as nodes distinguished by colors, and the relations between them are represented by the edges to form a network graph. phExplorer also allows the navigation through PTN to KEGG signaling pathways and protein-protein interaction network. Besides, phExplorer provides a homologous protein search function using BLAST (27), which helps to predict potential new or off-target proteins for the drug of interest. Because phExplorer was seamlessly integrated into PharmDB website, users can flexibly switch from text view to graphical network view with a single click of a button (Figure I-3A, I-3B, I-3C and I-3D). We also included exploratory functions to phExplorer so that users can continuously investigate their leads in the network view without switching back to the textual view to submit queries. It is not necessary to regenerate a whole new network graph because new nodes and edges are dynamically added to the present network graph if a query returns any results.

Data statistics of PharmDB

At present, PharmDB contains 11,860 proteins, 2,697 drugs and small molecules, and 688 diseases. Out of 11,860 proteins, 1,546 proteins are related to drug or disease. Among these proteins, 533, 403 and 610 proteins are linked to drug only, disease only and both, respectively (Figure I-4A). The rest of the proteins do not have any links to drugs or diseases although they are the part of the PPI and signal pathway data. In the case of drug, 2,697 drugs are linked to either or both of target proteins and diseases (Figure I-4B). Among them, 1155, 670 and 872 are linked to target only, disease only and both, respectively. 592 drugs are the FDA-approved while the rest 2105 chemicals are the experimental small molecules. PharmDB describes a total of 688 diseases. Among them, 166, 46 and 476 diseases are linked only to target, disease, and both, respectively (Figure I-4C).

Possible applications of PharmDB

PharmDB can be characterized as following. First, PharmDB provides a comprehensive and integrated database that contains knowledge of drug, disease and target protein on the framework of biological network that consists of PPI and signaling pathways. Thus, PharmDB alleviated the burden of aggregating and cleaning up heterogeneous data required for network-based analysis. Second, it provides a graphical browsing tool, phExplorer, which transforms assorted and multi-layered data into a graphical format. This network-based analysis of the relationships between drug, target and diseases may provide a predictive guide for the questions like novel therapeutic target identification, drug repositioning, potential side effect, and even designing optimal drug combination. For the suggestion of a new drug target, let's

assume that a drug of unknown target (DOUT) is known to treat four different diseases, and all of these diseases are commonly associated with one protein (target 3 in Figure I-5A left) although each of them may also have other associated protein as well. In this case, DOUT is likely to work to target 3 with high probability (Figure I-5A). The reliability of this connection would be proportional to the number of diseases that are shared with the particular target. PharmDB can be used to guide drug repositioning. Let's assume that a drug is known to work on disease 1 via its known target 1. PharmDB can find its homologous protein and fish out target 2 that is known to be associated with another disease. In this case, the drug can be applied to disease 3 via target 2. Moreover, if target 1 itself is also associated with another disease, this connection can also suggest the potential application of the drug to disease 2 (Figure I-5B). In a similar logic, a potential side effect of a drug can be deduced using PharmDB. A drug can bind to a protein homologous to its known target, which can give unwanted side effect. In addition, the target protein itself can generate a side effect due to its multiple activities. Likewise, PharmDB can be explored in numerous ways at users' own idea and interest.

Development of the shared neighborhood scoring algorithm

Integration of diseases, drugs, and proteins revealed numerous missing links between the components (Figure I-6A, dashed lines). However, whether these missing links resulted from “no actual connection” or “lack of knowledge” remains unclear. Before these missing links are experimentally validated, the probability of a link existing between two nodes can be logically predicted by evaluating the connections of their neighbors. We

hypothesized that if there are more shared nodes between two nodes, they are more probable to be connected. To test this, we compared the distributions of the count of shared nodes between connected and unconnected pairs. The result shows connected pairs have more shared nodes for all of the three kinds of link types (Figure I-7). Distribution comparisons of connected and unconnected pairs were conducted using Kolmogorov-Smirnov test. The test resulted in extremely low p-value($< 2.2\text{e-}16$) for the three kinds of pairs. It confirms that the number of shared nodes can be used as a score indicating the possibility between two nodes.

Based on this difference between connected and unconnected pairs, we designed a network topology discovery algorithm called “shared neighborhood scoring algorithm.” to score the likelihood of a link existing between two nodes of interest. This basic concept is similar to Swanson’s ABC model, which applies the transitivity rule to discover hidden knowledge from biomedical literature(28). In a pair of nodes (i and j) with no known connection between them, the likelihood of a connection depends on the number of commonly shared nodes comprised of directly linked nodes (m) and indirectly linked nodes (k and l) via additional nodes (Figure I-6B). In this algorithm, we first assigned a weight 1 to a direct link between the two nodes. For an indirect link or a virtual link, a weight is assigned by the connection probability. The connection probability was computed as the proportion of the number of directly connected pairs over the total number of all pairs between the two nodes for each case (Figure I-8). We then came up with the shared neighborhood score through the sum of the products of the weights of the two links bridging the two end nodes. As the SN score possesses a range of values

in each relation category (drug-protein, protein-disease, and drug-disease), we developed a normalization method using the connecting probability function of SN score distribution (Figure I-9). (see methods for details).

Prediction performance of the shared neighborhood score was tested using receiver operating characteristic (ROC) analysis(29) in the relationships among drug-protein, disease-protein, and drug-disease (Figure I-10A, B, and C, respectively). The test showed good discrimination for all three cases ($AUC > 0.7 \sim 0.9$). As there are risks that the SN score is overfitted to the PharmDB data, we checked robustness of the performance. This was measured by AUC changes in pharmacological networks generated through random selection or deletion of the pairs in the original data (Figure I-11, I-12). The result shows that the SN score is enough robust to random selection or deletion.

Development of a searching pipeline for drug repositioning

We tested whether the tripartite network map contained in PharmDB could be used to identify an unknown link between an existing drug and disease. To extract drug candidates that could be used to treat a disease of interest from PharmDB, we developed a workflow using Taverna Workbench(30) (<http://www.taverna.org.uk/>). In this process, PharmDB and an FDA-approved drug list were converted into the BioMart(31) format. The workflow started by retrieving the proteins linked to the disease of interest from PharmDB. Then, the drugs linked to the extracted proteins were retrieved (Figure I-13A). Among the extracted drugs, only FDA-approved drugs were selected. They were then subjected to e-Search provided by the

NCBI web service to look for previously published records for any connection between the drugs and disease of interest.

A case study for drug repositioning

As a case study, we chose squamous cell carcinoma (SCC) (MeSH descriptor: D002294), a subtype of lung cancer, and tested whether PharmDB could identify any drugs that have a potential for treating this type of cancer. A total of 11 drugs, all with no previous link to lung cancer, were extracted. We then evaluated the likelihood of the links between these drugs and SCC using the shared neighborhood algorithm (Figure I-13B).

Among the suggested drug candidates with high scores, we chose benzthiazide (TBZT) for further experimental validation since it is linked to a cancer-associated protein, carbonic anhydrase 9 (CA9) (Table I-2). TBZT is a kind of thiazide diurectic used for the treatment of high blood pressure and edema(32). PharmDB showed CA9 to be a commonly linked protein between TBZT and SCC (Figure I-14A). CA9, a carbonic anhydrase isoenzyme, is a transmembrane protein that plays an important role in pH regulation(33). The expression of CA9 is highly induced in various cancers under hypoxic conditions and is functionally important for the growth and survival of tumor cells(33). First, we selected squamous lung cancer cells (HCC-1588) and tested whether CA9 is actually induced in hypoxic conditions by Western blotting with its specific antibody. As expected, CA9 levels were significantly increased in hypoxic conditions (1% O₂) compared with those in normoxic conditions (20% O₂). We then administered different concentrations of TBZT to HCC-1588 cells under normoxic and hypoxic conditions; their effects on

cell proliferation were monitored by [^3H] thymidine incorporation. Acetazolamide (AZA), a known inhibitor of CAs, was also used as positive control(34). TBZT-suppressed cancer cell proliferation in a dose-dependent manner was possible only under hypoxic conditions, not normoxic conditions. In addition, its effective dose was much lower than that of AZA (Figure I-14C). The hypoxia-dependent cell death induced by TBZT was further confirmed by flow cytometry (Figure I-14D) and the activation of caspase 3 (Figure I-14E).

We tested whether TBZT could actually work as an inhibitor of CA9 and determined its exact inhibitory concentration using *in vitro* assay. As there are many CA9 isoforms with structural similarities, we used four different human CA isozymes (i.e., 1, 2, 9, and 12) for the assay. TBZT suppressed all of the four tested enzymes with similar K_i values (Figure I-14F). AZA also suppressed the activities of the four tested CAs, although the K_i values varied depending on the target enzymes. TBZT inhibited not only CA9 but also other CA isozymes with little discrimination, and thus its anti-proliferative activity might have resulted from its inhibitory activity against other CAs. However, among the CA isozymes, only CA9 and 12 are known to be induced in hypoxic conditions and have functional association with cancer(33). Thus, the efficacy of TBZT against HCC-1588 cells is likely to have resulted from its inhibition of CA9 and possibly CA12. To confirm this possibility further, we tested whether the forced expression of CA9 would compensate for the anti-proliferative activity of CA9 under hypoxic conditions. Cell proliferation was reduced to 20% of the control cells upon TBZT treatment, but the exogenous supplementation of CA recovered the proliferation by up to 90% (Figure I-

14H). This result validates that the anti-proliferative activity of TBZT against HCC-1588 cells mainly involves CA9. Even if further chemical optimization of TBZT is required to improve efficacy and specificity, these results suggest a possible application of TBZT for further development against lung cancer via CA9.

CONCLUTIONS

In summary, PharmDB is a novel database that shows drugs, targets, diseases and their relationships in an integrated fashion. The data were gathered and integrated using unique identifiers: GeneID, PubChem ID, MeSH descriptor for targets, drugs, and diseases, respectively. In this database, we focused to human endogenous therapeutic targets. PharmDB provides not only a simple retrieval of database entry but also a sub-network retrieval of the global map using graphical user interface. This study also demonstrates that drug repositioning can be rapidly guided by a knowledge platform PharmDB, which is a tripartite pharmacological network database and the “Shared neighborhood scoring”(SNS) algorithm which is a network topology discovery algorithm. Although the information contained in this database can be obtained through manual network navigation, the entire procedure can be automated through the BioMart and Taverna Workbench to produce a more convenient and flexible workflow. Aside from drug repositioning, the network map of PharmDB can also be applied to other purposes, such as the prediction of drug mode-of-action, off-target effects, and even the design of optimal drug combinations for a disease of interest.

METHODS

Construction of PharmDB

To integrate the data in the existing databases that contain different identifiers, we assigned the following standard identifiers (IDs): PubChem CID for drug, GeneID for protein, and MeSH descriptor for disease. We denoted drugs that do not have a PubChem CID as PubChem SID and added a minus sign to distinguish them from PubChem CID. For tagging the standard IDs, we downloaded the reference databases from the websites of PubChem (<ftp://ftp.ncbi.nlm.nih.gov/pubchem/>), EntrezGene (<ftp://ftp.ncbi.nlm.nih.gov/gene/>), and MeSH (<http://www.nlm.nih.gov/mesh/filelist.html>). We then parsed the data files and generated cross-reference tables.

We constructed a comprehensive drug-protein-disease tripartite network by integrating the link information from the seven databases, namely, TTD(14) (http://bidd.nus.edu.sg/group/cjttd/TTD_ns.asp), DrugBank(15) (<http://redpoll.pharmacy.ualberta.ca/drugbank/>), ChemBank(16) (<http://chembank.broad.harvard.edu/>), PharmGKB(35) (<http://www.pharmgkb.org/>), KEGG(17) (<http://www.genome.jp/kegg/>), OMIM(36) (<http://www.ncbi.nlm.nih.gov/omim>), and GAD(37) (<http://geneticassociationdb.nih.gov/>) (Table I-1 for detailed information). As the existing databases have their own unique ID systems, we tagged the standard IDs using an in-house script. For the entities that were not tagged by the in-house script, we manually assigned them with appropriate IDs.

Implemented software for PharmDB

We implemented PharmDB using MySQL which contains the tables describing drug, target and disease as nodes, and another set of tables including the links between them. The target tables contain not only therapeutic proteins, but also the proteins that are only involved in protein-protein interactions and signaling pathways without any linkages to drugs or diseases. There are four tables containing the information of the links between drug-target, target-disease, disease-drug and target-target. Lastly, there is a reference table that contains the PubMed references. These references are used as the evidences supporting the relatedness among the entities. The web interface of PharmDB is serviced using Apache Web server. All of the scripts for the web interface were written in Perl. For the graphical view and navigation of network figures, Java applet-based network viewer named phExplorer was developed. phExplorer was implemented using Prefuse graphical toolkit (38).

Shared neighborhood scoring algorithm

The shared neighborhood scoring algorithm is based on the basic principle that the probability that a connection exists between two nodes (i and j) is proportional to the number of nodes commonly shared between the original two nodes, i and j (Figure I-6B). The shared neighborhood score S_{ij} is defined as $S_{ij} = \sum_k w_{ik}w_{kj}$. In this equation, i and j indicate the indices of a pair of nodes; k is the index of a shared neighbor node; and w_{ik} is the weight of a link between i and k . The link between i (or j) and k can be real or virtual (i.e., having no known connection but is expected to be connected).

Thus, we can define w_{ik} as $w_{ik} = a_{ik} + P(S_{ik}^{(0)})\delta_{a_{ik},0}$. Here, $a_{ik} = 1$ and $\delta_{a_{ik},0} = 0$ if the link between i and k is real; and $a_{ik} = 0$ and $\delta_{a_{ik},0} = 1$ if the link is virtual. When there are only direct connections between node i and node j , a 0th-order shared neighborhood score $S_{ij}^{(0)}$ becomes $S_{ij}^{(0)} = \sum_k a_{ik}a_{kj} = n$ (0, 1, 2, 3 ...), where n is the number of bridging nodes between node i and node j . $P(S_{ik}^{(0)})$ is a connection probability that depends on the value of the 0th-order shared neighborhood score $S_{ik}^{(0)}$. For the 0th-order shared neighborhood score $S_{ik}^{(0)} = n$ ($n = 0, 1, 2, \dots$), the function $P(n)$ is defined as follows:

$$P(n) = \frac{(\text{number of connected pairs})}{(\text{number of pairs})} \text{ for } n = 1, 2, \dots$$

$$P(n) = 0 \text{ for } n = 0$$

Based on the probability above, non-linear regression was carried out to extract connecting probability functions. Logistic function was used for this (Figure I-8).

The shared neighborhood score S_{ij} then becomes $S_{ij} = S_{ij}^{(0)} + S_{ij}^{(1)} + S_{ij}^{(2)} = \sum_k (a_{ik} + P(S_{ik}^{(0)})\delta_{a_{ik},0}) (a_{kj} + P(S_{kj}^{(0)})\delta_{a_{kj},0})$, where $S_{ij}^{(0)} = \sum_k a_{ik}a_{kj}$, $S_{ij}^{(1)} = \sum_k (a_{ik}P(S_{kj}^{(0)})\delta_{a_{kj},0} + a_{kj}P(S_{ik}^{(0)})\delta_{a_{ik},0})$, and $S_{ij}^{(2)} = P(S_{ik}^{(0)})\delta_{a_{ik},0}P(S_{kj}^{(0)})\delta_{a_{kj},0}$. Here, the 1st-order term $S_{ij}^{(1)}$ is added when some nodes are linked directly to node i (or j) but linked indirectly to node j (or i). The 2nd-order term $S_{ij}^{(2)}$ is considered only when some nodes are linked indirectly to both node i and node j . In Figure II-1C, because node

m is the only shared neighbor of node i and node j , $S_{ij}^{(0)} = 1$. To obtain the 1st-order shared neighborhood score $S_{ij}^{(1)}$ or the 2nd-order shared neighborhood score $S_{ij}^{(2)}$, connection probability $P(S_{ik}^{(0)})$ is calculated beforehand. As a pair (i, k) is mediated by two nodes, $w_{ik} = P(2)$, Path (i, k, j) is composed of an indirect link (i, k) and a direct link (k, j) . Similarly, $w_{il} = P(3)$ and $w_{lj} = P(1)$. Path (i, l, j) is composed of both indirect links (i, l) and (l, j) . The total shared neighborhood score is thus $S_{ij} = S_{ij}^{(0)} + S_{ij}^{(1)} + S_{ij}^{(2)} = 1 + P(2) + P(3)P(1)$. When calculating $S_{ij}^{(2)}$, we omitted a link between node i and j to remove the dependency of the measure on the existence of a link between i and j , which is the so-called “leave-one-out approach”.(39)

The shared neighborhood score is proportional to the number of shared neighborhood nodes. So there is no an upper limit on score. The problem is that the amount of data is not evenly distributed on each relation category. So, even if two different types of relations have identical score, their connecting possibility can't be regarded as identical. For that reason, we normalized the shared neighborhood score using connecting probability function (Figure I-9).

ROC analysis

We generated the ROC curves and calculated the areas under the ROC curves (AUC) using the Bioconductor ROC library (1.10.0). For the test, we used the original data set as the gold standard set. The ROC test set was restricted to non-zero scores because over 99% of the zero-score pairs are

unconnected. If we took all possible pairs, including zero-score pairs, to be the test set, a large true negative would give a specificity of ~1.

Procedure for extracting drug repositioning candidates

We used the Taverna workflow system to extract drug candidates that show potential in the treatment of SCC. We converted three tables (i.e., link table, node table, and synonym table) of PharmDB into the BioMart format to allow the data to be used for the Taverna Workbench (PharmDB BioMart services are now available on <http://biomart.i-pharm.org/>). The link table of PharmDB consists of three relationship categories, namely, drug-protein, protein-disease, and drug-disease. The workflow for extracting drug candidates for SCC was established using the following steps. First, we submitted the MeSH descriptor of SCC (D002294) as an input and retrieved the proteins linked to SCC from the PharmDB link table using the relationship category “disease-protein.” The extracted proteins were then submitted to the PharmDB link table under the relationship category of “drug-protein”; the drugs linked to the submitted proteins were then obtained. The obtained drugs were converted to generic names using the node table of PharmDB. Among the retrieved drugs, we selected only FDA-approved drugs and subjected them to e-Search provided by the NCBI web service to determine whether there were any previous reports demonstrating the linkage between the selected drugs and SCC. Finally, we selected the drugs that have not previously been applied to SCC as therapeutic agents.

Cell culture and materials

The HCC-1588 cell line was obtained from the Korean cell line bank and was maintained in RPMI (Hyclone) containing 10% fetal bovine serum and 1% antibiotics. Antibody against caspase-3 and tublin (Cell Signaling Technology) were purchased. M73 monoclonal antibody to CA9 was obtained from Dr. S. Pastorekova (Slovak Academy of Science, Slovak Republic). TBZT and AZA were purchased from Sigma.

Thymidine incorporation assay

To determine the effect of TBZT on cell proliferation, HCC-1588 cells were treated with TBZT in 2% serum-containing media for 48 h under normoxic (20% O₂) and hypoxic (1% O₂) conditions. AZA was used as positive control. pcDNA3-CA9 vector and empty vector (Dr. J.-Y. Kim, National Cancer Center, Korea) were transfected into HCC-1588 cells using Lipofectamine 2000 (Invitrogen). After 24 h incubation, TBZT was added to 2% serum-containing media for 48 h under hypoxic conditions. [³H] thymidine at 1 µCi/ml was added to the culture medium and was incubated for 4 h. The incorporated thymidine was measured by liquid scintillation counter (Wallac).

Flow cytometry

HCC-1588 cells were treated with TBZT (0.4, 2, 10 M) in 2% serum-containing medium for 48 h under normoxic and hypoxic conditions. AZA was used as positive control. The treated cells were fixed with 70% ethanol for 1 h at 4°C, washed twice with ice-cold PBS, and stained with propidium iodide (50 µg/ml) containing 0.1% sodium citrate, 0.3% NP-40 (nonylphenoxypolyethoxyethanol 40), and 50 µg/ml RNase A for 40 min.

The cells were subjected to flow cytometry (FACSCalibur, Becton-Dickinson) to evaluate the apoptotic cells by counting the sub-G1 cells. For each sample, 20,000 cells were analyzed using Cell Quest Pro software.

Enzyme activity

An applied photophysics stopped-flow instrument was used for assaying CA-catalyzed CO₂ hydration activity(40). Following the initial rates of the CA-catalyzed CO₂ hydration reaction for a period of 10–100 s, phenol red (at a concentration of 0.2 mM) was used as the indicator, working at the absorbance maximum of 557 nm in 20 mM HEPES buffer (pH 7.5) and 20 mM Na₂SO₄ (to maintain the constant ionic strength). For the determination of the kinetic parameters and inhibition constants, the CO₂ concentrations used ranged from 1.7–17 mM. For each inhibitor, at least six traces of the initial 5–10% of the reaction were used for determining the initial velocity. The uncatalyzed rates were determined in the same manner and were subtracted from the total observed rates. Stock solutions of the inhibitor (0.1 mM) were prepared in distilled-deionized water and diluted to 0.01 nM with distilled-deionized water. Inhibitor and enzyme solutions were preincubated together for 15 min–72 h at room temperature (15 min) or 4°C (all other incubation times) prior to assay to allow the formation of the enzyme-inhibitor complex or the eventual active site mediated hydrolysis of the inhibitor. The inhibition constants were obtained by non-linear least-squares methods using PRISM 3 as previously described. The mean values were represented from at least three different determinations.(33,41)

ABBREVIATIONS

MeSH: Medical Subject Headings

ICD: International Classification of Diseases

PTN: pharmacological tripartite network

PPI: protein-protein interaction

BIND: Biomolecular INteraction Network Database

HPRD: Human Protein Reference Database

DIP: the Database of Interacting Proteins

MINT: the Molecular INteraction database

TTD: Therapeutic Target Database

PharmGKB: The Pharmacogenomics Knowledge Base

KEGG: Kyoto Encyclopedia of Genes and Genomes

OMIM: Online Mendelian Inheritance in Man

GAD: Genetic Association Database

DOUT: drug of unknown target

SNS: shared neighborhood scoring

SN score: shared neighborhood score

ROC: receiver operating characteristic

AUC: area under curve

AZA: Acetazolamide

CA: carbonic anhydrase

SCC: squamous cell carcinoma

TBZT: Thia-benzthiazide

KS test: Kolmogorov-Smirnov test

REFERENCES

1. Ashburn, T.T. and Thor, K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov*, **3**, 673-683.
2. Newman, M.E.J., Strogatz, S.H. and Watts, D.J. (2001) Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, **6402**, -.
3. Strogatz, S.H. (2001) Exploring complex networks. *Nature*, **410**, 268-276.
4. Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat Rev Drug Discov*, **4**, 45-58.
5. Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J.C., Hernandez-Boussard, T., Rees, C.A., Cherry, J.M., Botstein, D., Brown, P.O. *et al.* (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res*, **31**, 219-223.
6. Olson, A.J., Tully, T. and Sachidanandam, R. (2005) GeneSeer: a sage for gene names and genomic resources. *BMC Genomics*, **6**, 134.
7. Pearson, H. (2001) Biology's name game. *Nature*, **411**, 631-632.
8. Butcher, E.C., Berg, E.L. and Kunkel, E.J. (2004) Systems biology in drug discovery. *Nat Biotechnol*, **22**, 1253-1259.
9. Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, **33**, D54-58.

10. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **35**, D5-12.
11. Chen, J., Swamidass, S.J., Dou, Y., Bruand, J. and Baldi, P. (2005) ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics*, **21**, 4133-4139.
12. Lipscomb, C.E. (2000) Medical Subject Headings (MeSH). *Bull Med Libr Assoc*, **88**, 265-266.
13. World Health Organization. (2005). 10th revision, 2nd ed. World Health Organization, Geneva, pp. 1 CD-ROM.
14. Chen, X., Ji, Z.L. and Chen, Y.Z. (2002) TTD: Therapeutic Target Database. *Nucleic Acids Res*, **30**, 412-415.
15. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. and Woolsey, J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, **34**, D668-672.
16. Strausberg, R.L. and Schreiber, S.L. (2003) From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science*, **300**, 294-295.
17. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34**, D354-357.
18. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. and

- Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, **30**, 303-305.
19. Lim, J., Hao, T., Shaw, C., Patel, A.J., Szabo, G., Rual, J.F., Fisk, C.J., Li, N., Smolyar, A., Hill, D.E. *et al.* (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, **125**, 801-814.
 20. Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173-1178.
 21. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957-968.
 22. Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A. and Cesareni, G. (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, **6 Suppl 4**, S21.
 23. Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. and Chinnaiyan, A.M. (2005) Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, **23**, 951-959.
 24. Apic, G., Ignjatovic, T., Boyer, S. and Russell, R.B. (2005) Illuminating drug discovery with biological pathways. *FEBS Lett*, **579**,

1872-1877.

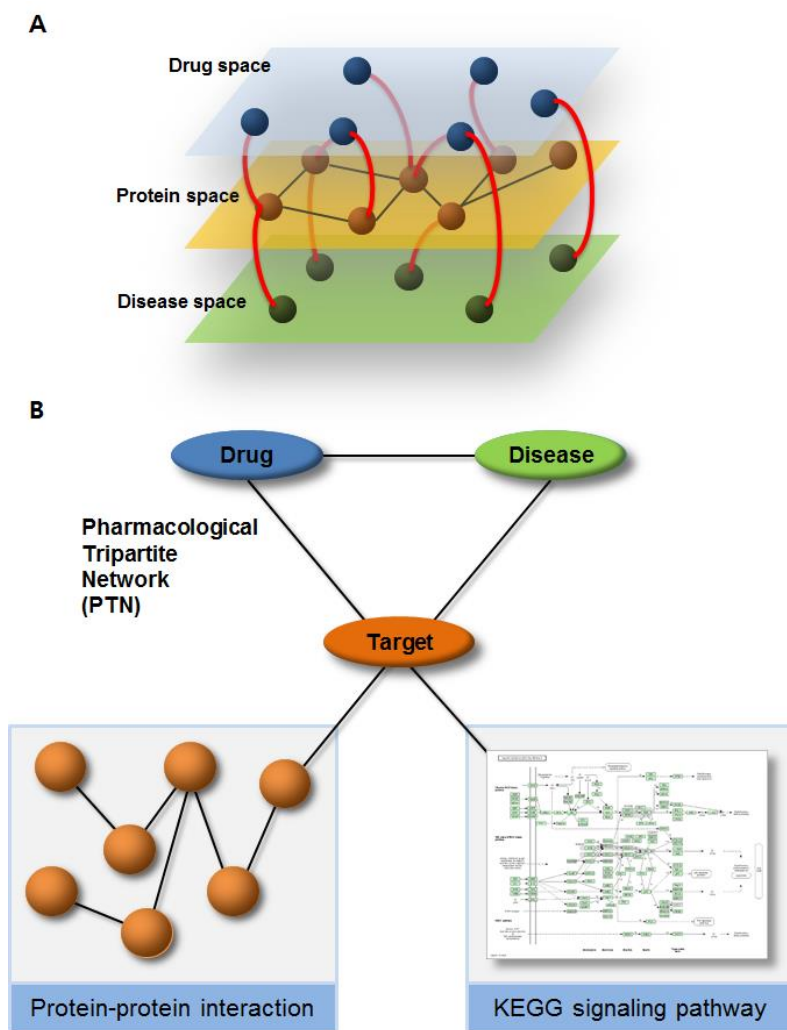
25. Breitkreutz, B.J., Stark, C. and Tyers, M. (2003) Osprey: a network visualization system. *Genome Biol*, **4**, R22.
26. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498-2504.
27. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.
28. Swanson, D.R. and Smalheiser, N.R. (1997) An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, **91**, 183-203.
29. Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861-874.
30. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, **34**, W729-732.
31. Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. and Kasprzyk, A. (2009) BioMart--biological queries made easy. *BMC Genomics*, **10**, 22.
32. Havard, C.W. and Wood, P.H. (1960) Clinical evaluation of benzthiazide, an oral diuretic. *Br Med J*, **1**, 1773-1776.
33. Supuran, C.T. (2008) Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nat Rev Drug Discov*, **7**, 168-181.

34. Xiang, Y., Ma, B., Li, T., Yu, H.M. and Li, X.J. (2002) Acetazolamide suppresses tumor metastasis and related protein expression in mice bearing Lewis lung carcinoma. *Acta Pharmacol Sin*, **23**, 745-751.
35. Klein, T.E., Chang, J.T., Cho, M.K., Easton, K.L., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D.E. *et al.* (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J*, **1**, 167-170.
36. Online Mendelian Inheritance in Man, O.T.M.-N.I.o.G.M., Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 03/03/2009. World Wide Web URL: <http://www.ncbi.nlm.nih.gov/omim/>.
37. Becker, K.G., Barnes, K.C., Bright, T.J. and Wang, S.A. (2004) The genetic association database. *Nat Genet*, **36**, 431-432.
38. Heer, J., Card, S.K. and Landay, J.A. (2005), *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, Portland, Oregon, USA.
39. Goldberg, D.S. and Roth, F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A*, **100**, 4372-4376.
40. Khalifah, R.G. (1971) The carbon dioxide hydration activity of carbonic anhydrase. I. Stop-flow kinetic studies on the native human isoenzymes B and C. *J Biol Chem*, **246**, 2561-2573.
41. Maresca, A., Temperini, C., Vu, H., Pham, N.B., Poulsen, S.A.,

Scozzafava, A., Quinn, R.J. and Supuran, C.T. (2009) Non-Zinc Mediated Inhibition of Carbonic Anhydrases: Coumarins Are a New Class of Suicide Inhibitors. *J Am Chem Soc*, **131**, 3057-3062

FIGURES AND TABLES

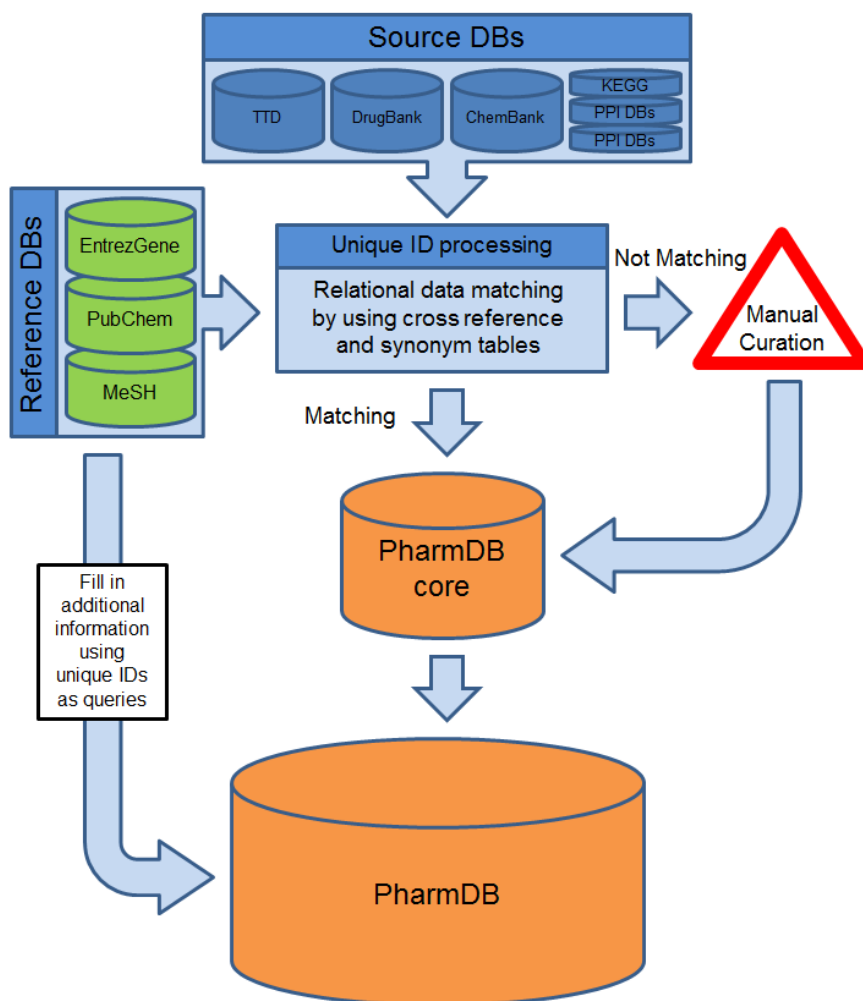
Figure I-1. Data model and architecture of-PharmDB



(A) Illustration of simplified data model of drug discovery, which was implemented in PharmDB. PharmDB contains the information of drug, target and disease and their interconnections. (B) The database contains the

information of commercially available and experimental drugs, therapeutic target proteins and diseases and their relationships in the format of pharmacological tripartite network (PTN). The therapeutic proteins are also linked to comprehensive biological network consisting of protein-protein interaction (PPI) and KEGG signaling pathways.

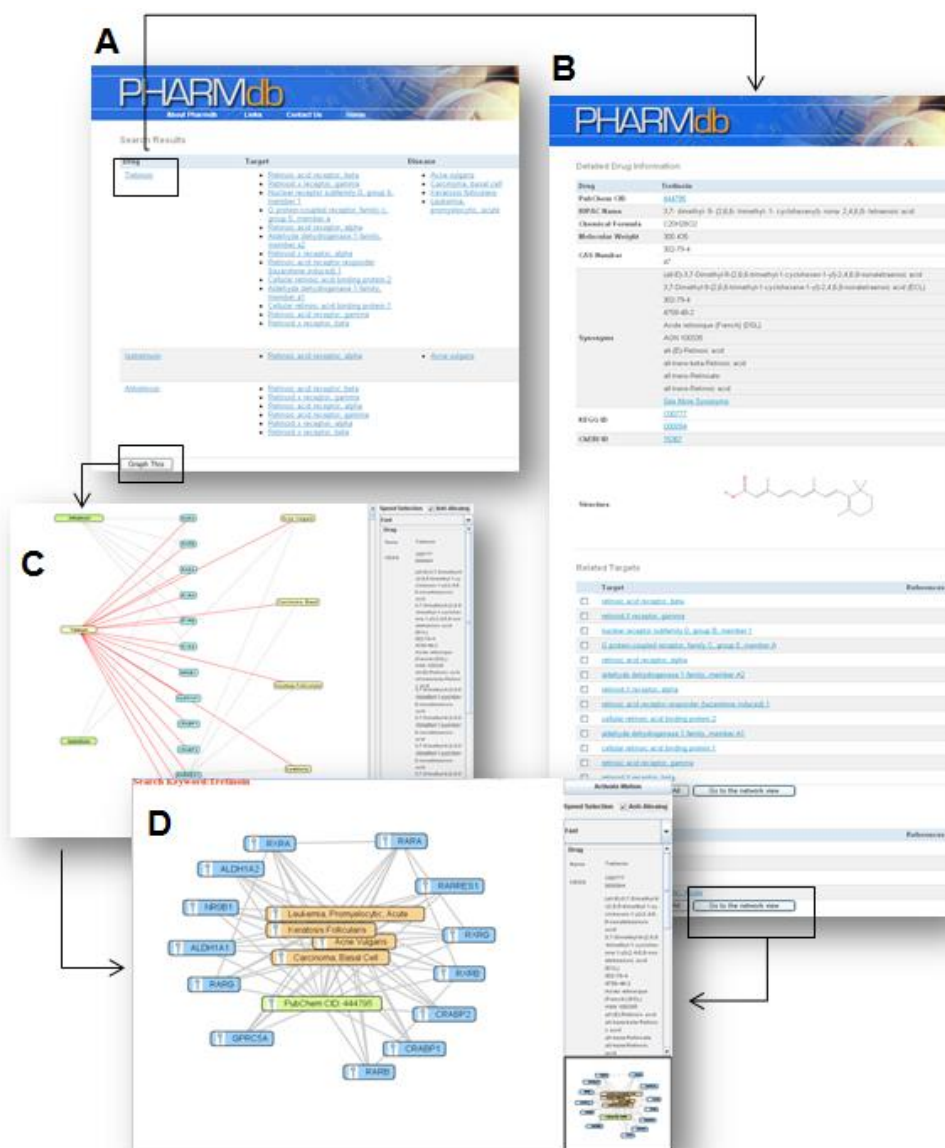
Figure I-2. Data integration process of PharmDB



The publicly available drug, target and disease databases along with the reference databases were imported to our MySQL database. We then conducted data cleanup and integration process. GeneID, PubChem CID and SID, and MeSH terms were assigned to each entry of drug, target and disease, respectively, utilizing cross reference and synonym tables provided by both

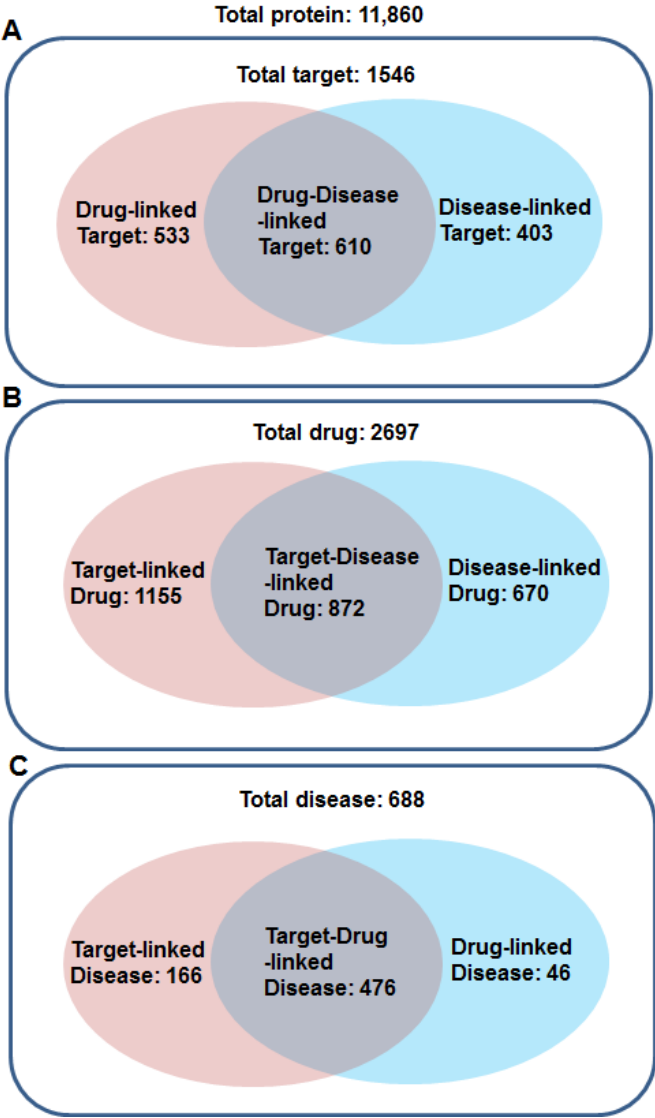
original and reference databases. The entries lacking any cross reference or synonym information were manually curated. Upon the completion of unique identifier assignment, the entries with proper identifiers formed PharmDB core. We then added feature descriptions, such as cross database identifiers and synonyms, obtained from reference databases into each entry.

Figure I-3. Different output patterns in PharmDB



(A) The initial search results of a sample query, tretinoin, the acid form of vitamin A that is used to treat acne. (B) Detailed text view of tretinoin is generated by the click of tretinoin link. The output gives the information on PubChem CID, IUPAC name, Chemical Formula, Molecular weight, CAS number, synonyms, KEGG and ChEBI IDs and chemical structure. In addition, it provides the lists of targets and diseases that are linked to tretinoin. (C) Radial network graph layout of the phExplorer showing proteins and disease linked to tretinoin. The network map can be displayed according to the selection of targets and diseases of interest. (D) The tripartite network view of the search result. The initial output of the query can be converted to this type of view by clicking “graph this” button at the bottom. The tripartite network view can be also converted to radial network view by selecting the entities of interest and clicking “Explorer” button.

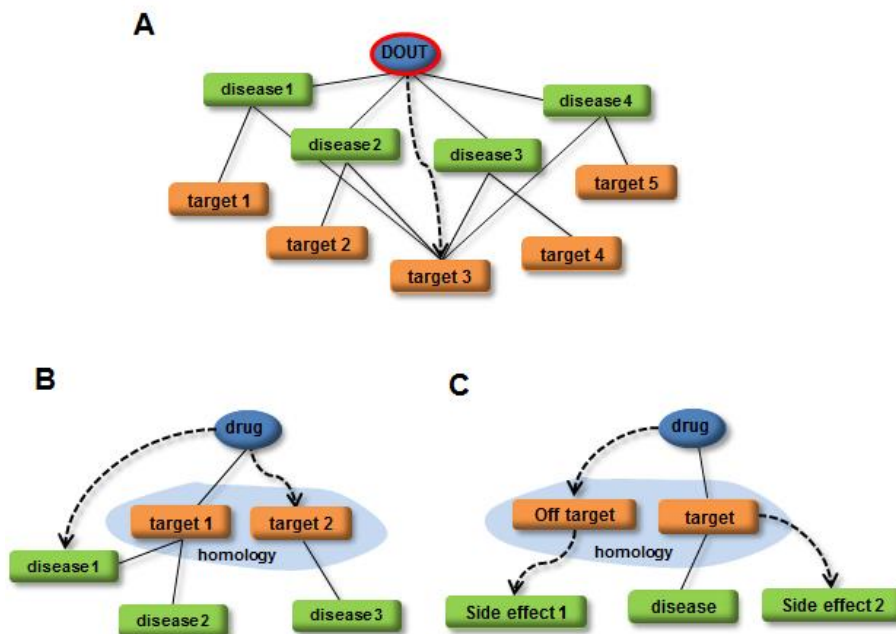
Figure I-4. Data statistics of PharmDB



(A) Diagram of target data. PharmDB has 10,580 proteins, of which 1,546 proteins are linked to either or both of drug and disease. Among these proteins, 533, 403 and 610 proteins are linked to drugs, diseases and both, respectively.

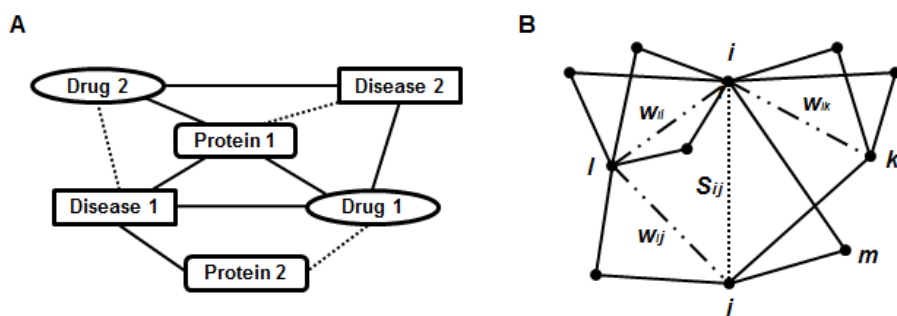
The rest of the proteins do not have any links to drugs or diseases. (B) Diagram of drug data. PharmDB contains the information of 2,697 drugs and bioactive chemicals. Among these 1,155 and 670 drugs are linked only to target and disease, respectively, and 872 drugs are linked to both of target and disease. (C) Diagram of disease data. PharmDB contains 688 human diseases excluding infectious diseases. 166 and 46 diseases are linked only to target and drug, respectively, and 476 diseases are linked both to target and drug.

Figure I-5. Possible applications of PharmDB



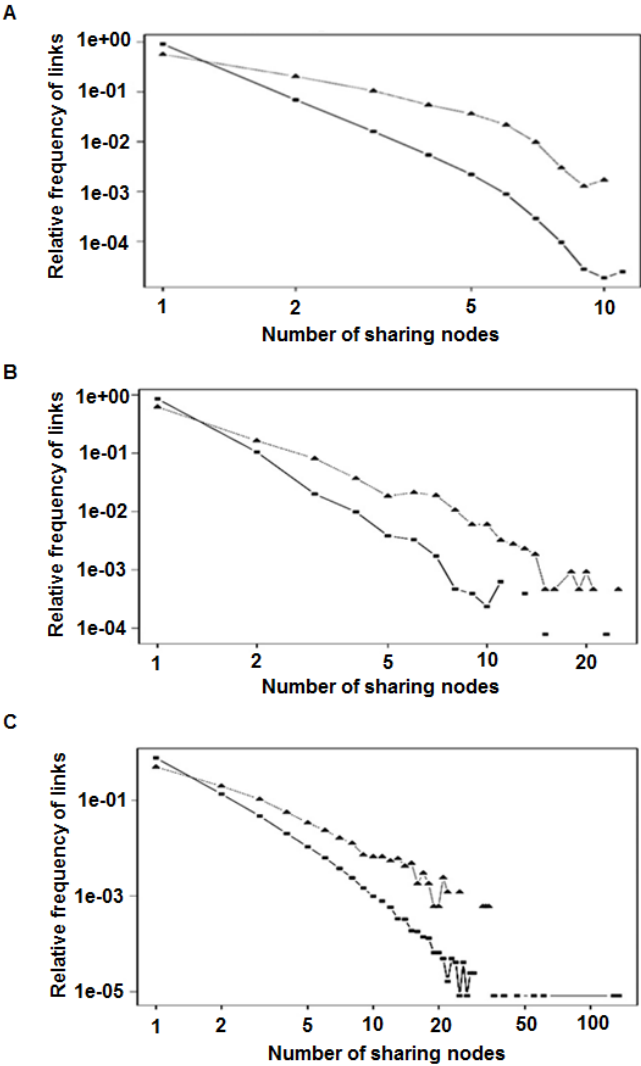
(A) A drug of unknown target (DOUT) is known to treat four diseases. If the four diseases are commonly associated with target 3, it can be the likely target of DOUT. (B) A drug is known to treat disease 1 via target 1. If target 1 is known to be also associated with disease 2, this drug can also work to disease 2. PharmDB also searches a protein, target 2, that is homologous to target 1. If target 2 is associated with disease 3, the drug can be also applied for the new indication to treat disease 3. (C) Homology search of a target protein can be also used to predict potential side effect. The information around the target protein can give the insight into a potential side effect.

Figure I-6. Missing links in the PharmDB network



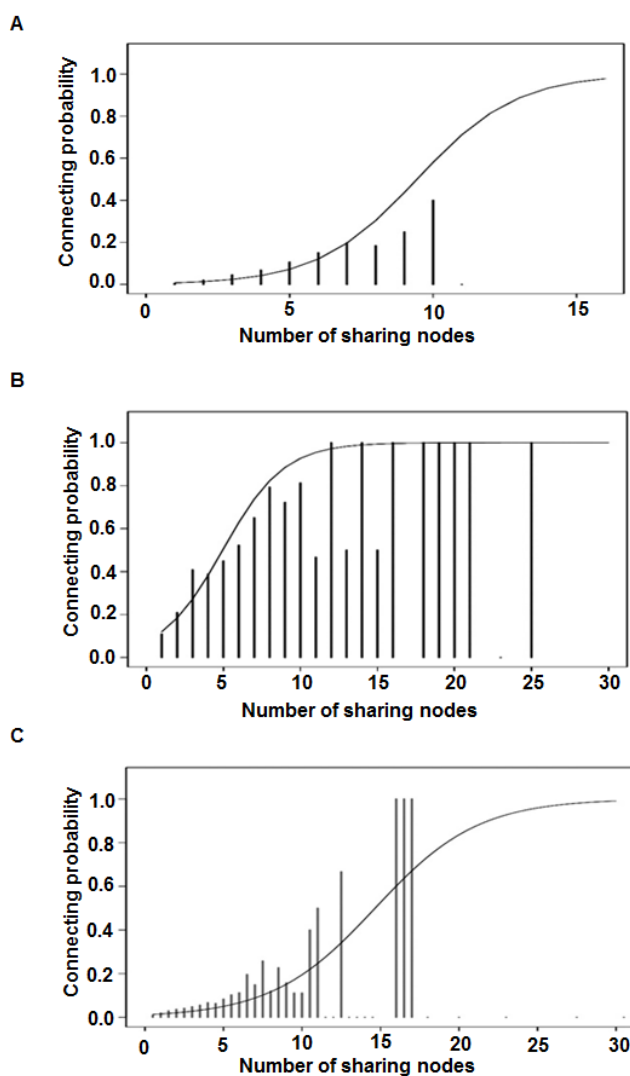
(A) Triangular circuit among diseases, proteins, and drugs with missing links (dotted lines). (B) Schematic representation of the shared neighborhood scoring algorithm calculating the likelihood of the existence of a particular link (solid line: real link; double-dotted dashed line: virtual link; dotted line: missing link).

Figure I-7. Shared neighborhood node distribution



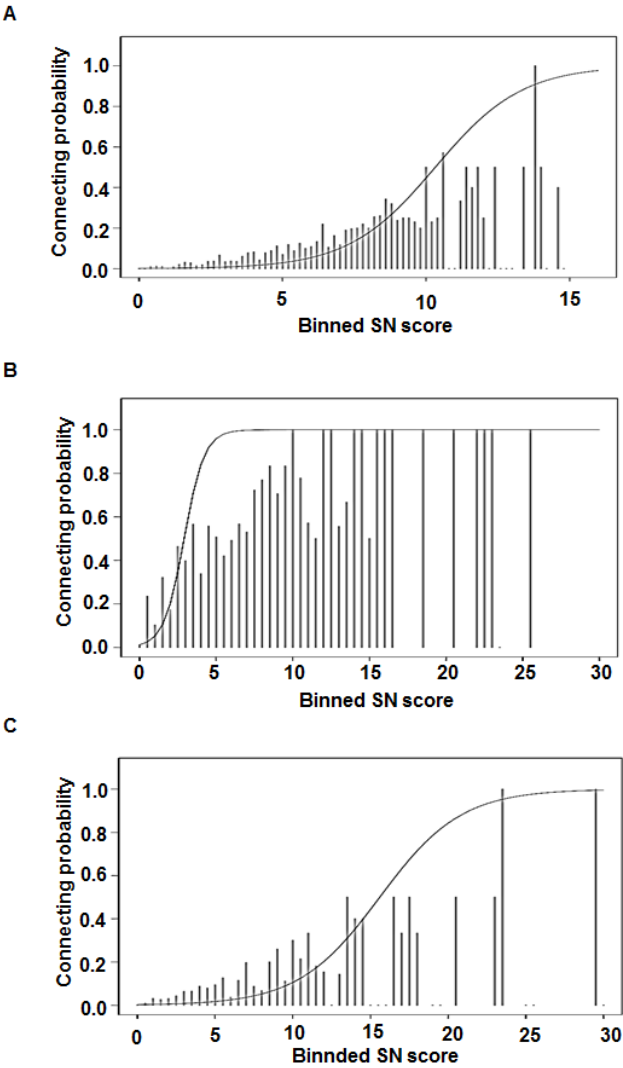
Shared neighborhood node distribution comparison between connected links and unconnected links in drug-protein relation (A), disease-protein relation (B) and drug-disease relation (C)

Figure I-8. Non-linear regression results for extracting connecting probability function



Non-linear connecting probability model based on the number of nodes commonly shared between two nodes. (A) drug-protein. (B) disease-protein. (C) drug-disease.

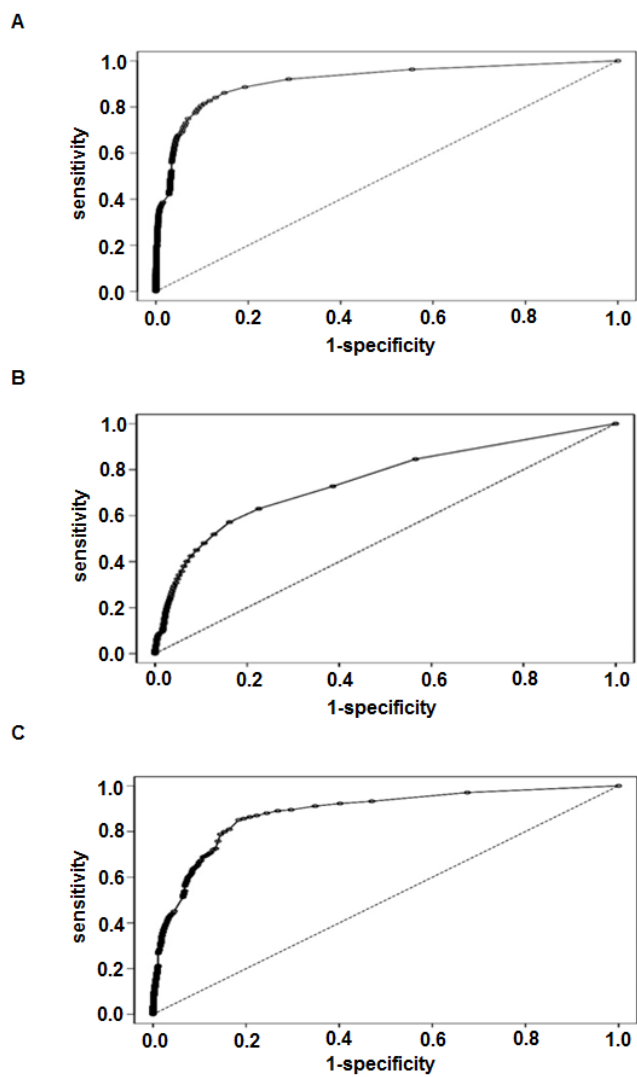
Figure I-9. Connecting probability model for SN score



The shared neighborhood score is proportional to the number of shared neighborhood nodes. As the amount of data is not evenly distributed on each relation category, even if two different types of relations have identical score, their connecting possibility can't be regarded as identical.

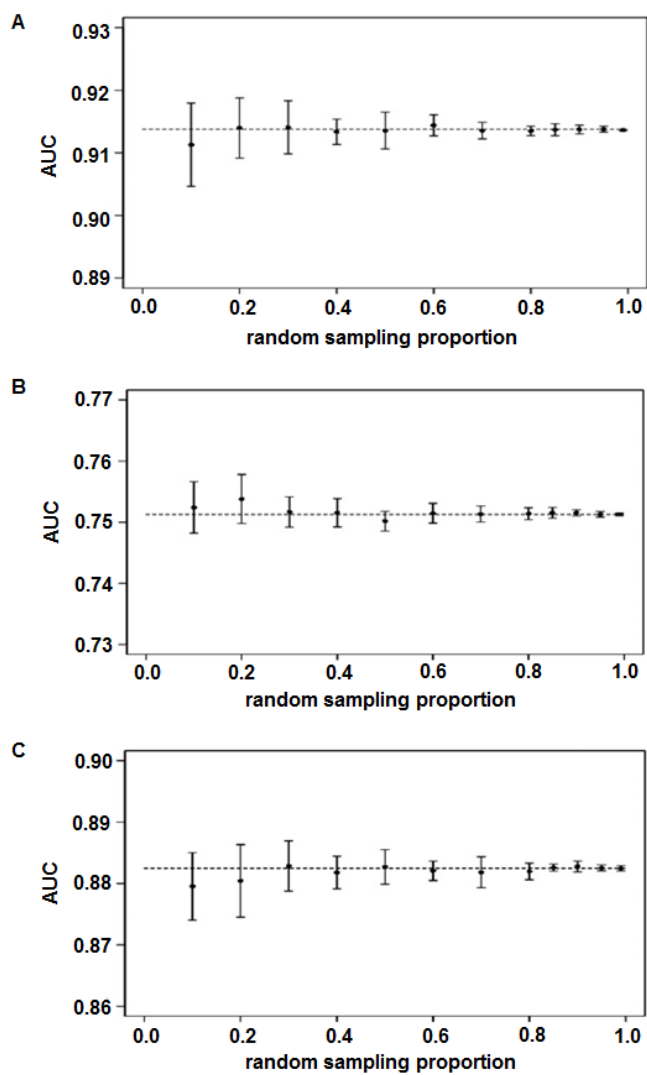
Therefore the shared neighborhood score is normalized using connecting probability function. (A) drug-protein (bin size: 0.2). (B) disease-protein (bin size: 0.5). (C) drug-disease (bin size: 1).

Figure I-10. ROC curves of the shared neighborhood scoring algorithm



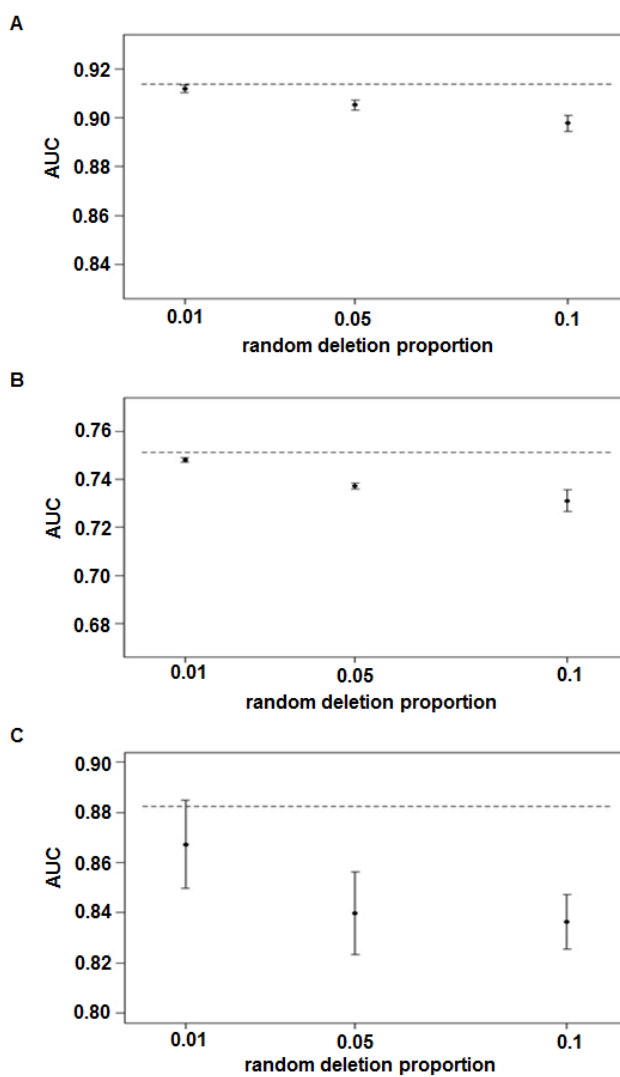
(A) For drug-protein relation, AUC = 0.914. (B) For disease-protein relation, AUC = 0.751. (C) For drug-disease relation, AUC = 0.882

Figure I-11. AUC changes by random sampling



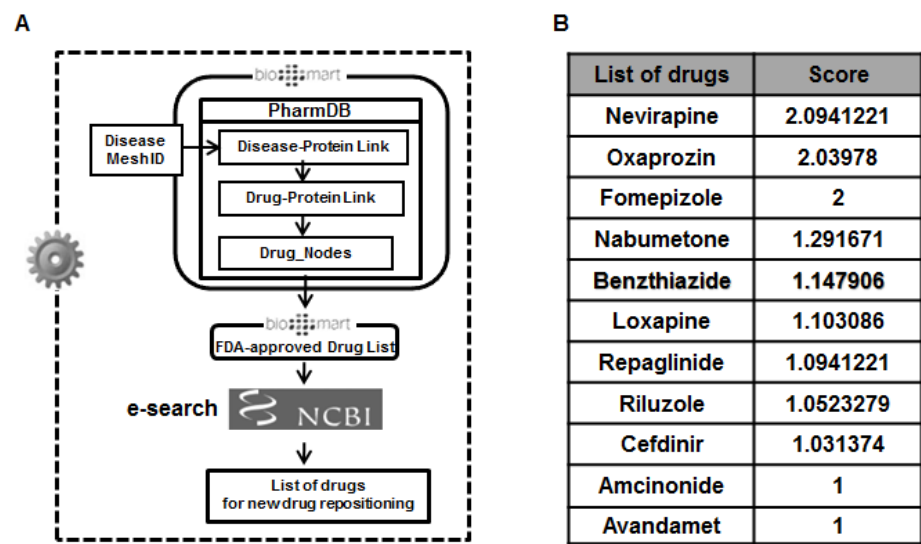
(A) drug-protein relation. (B) disease-protein relation. (C) drug-disease relation.

Figure I-12. AUC changes by random deletion



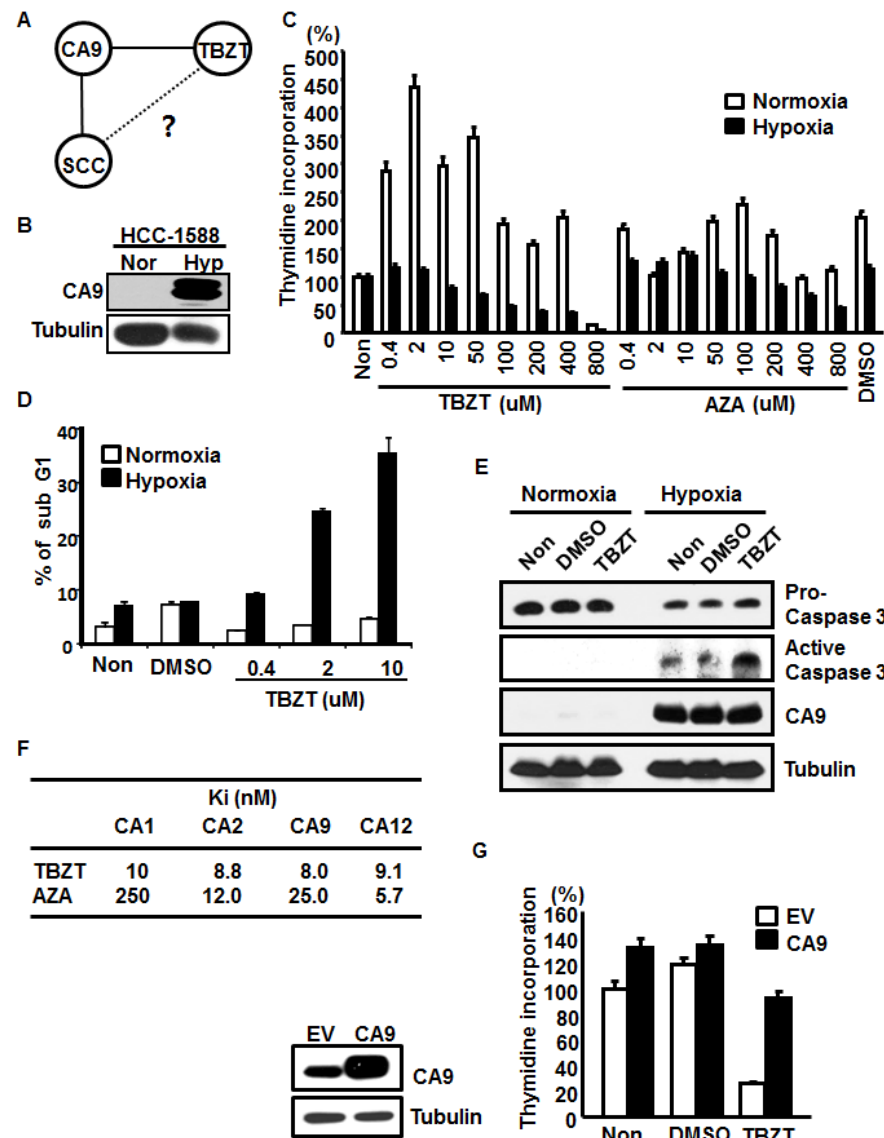
(A) drug-protein relation. (B) disease-protein relation. (C) drug-disease relation.

Figure I-13. A workflow for drug repositioning



(A) Schematic representation of the automatic workflow for drug repositioning. (B) List of drug candidates that show potential application in squamous cell lung carcinoma with scores resulting from the shared neighborhood scoring algorithm.

Figure I-14. Experimental validation of the effect of TBZT against SCC



(A) PharmDB showing the connections between SCC and carbonyl anhydrase (CA9) and between CA9 and TBZT, but not the connection between TBZT and SCC. (B) Cellular levels of CA9 in the SCC cell line, HCC-1588, under

normoxic and hypoxic conditions. (C) Antiproliferative activity of TBZT was monitored by [³H] thymidine incorporation under normoxic and hypoxic conditions. AZA was used as positive control. The effect of TBZT on cell death was monitored by counting sub-G1 cells (D) and caspase-3 activation (E). (F) *In vitro* inhibition of TBZT and AZA against CA isoforms (i.e., 1, 2, 9, and 12). (G) HCC-1588 cells, transfected with an empty vector (EV) or CA9, were treated with TBZT under normoxic and hypoxic conditions. Cell proliferation was monitored as above.

Table I-1. Summary of PharmDB data sources

	Drug-Protein Relation	Protein-Disease Relation	Drug-Disease Relation	DB version / Download date
TTD	O	O	O	09/05/2006
DrugBank	O		O	Version 2.5
PharmGKB	O		O	04/02/2009
KEGG	O			03/02/2009
OMIM		O		03/03/2009
GAD		O		01/10/2007
ChemBank			O	03/07/2009

Each database provides the linkage information of the indicated pairs among drugs, proteins, and diseases.

Table I-2. List of drug candidates and drug-associated proteins

Drugs	Drug-associated Protein		ATC code
		*PK and PD	
Nevirapine		ABCB1(ATP-binding cassette, sub-family B, member 1), CYP2B6(cytochrome P450, family 2, subfamily B, polypeptide 6), CYP3A5 (cytochrome P450, family 3, subfamily A, polypeptide 5)	J05AG01
Oxaprozin		PTGS2(prostaglandin-endo peroxide synthase 2)	M01AE12
Fomepizole		ADH1A(Alcohol dehydrogenase 1A), ADH1B(Alcohol dehydrogenase 1B), ADH1C(Alcohol dehydrogenase 1C), CAT (catalase)	V03AB34
Nabumetone		PTGS1(prostaglandin-endo peroxide synthase 1), PTGS2(prostaglandin-endo peroxide synthase 2)	M01AX01
Benzthiazide	CA1(Carbonic anhydrase1), CA2(Carbonic anhydrase2), CA4(Carbonic anhydrase4), CA9(Carbonic anhydrase9), KCNMA1(potassium large conductance calcium-activated channel, subfamily M, alpha member 1), SLC12A3(solute carrier family 12, member 3)		
Loxapine		CYP2D6 (cytochrome P450, family2, subfamilyD, polypeptide6), DRD1(dopamine receptor D1), DRD2(dopamine receptor D2)	N05AH01
Repaglinide	KCNJ1(potassium inwardly-rectifying channel, subfamily J, member 1)	ABCC8(ATP-binding cassette, sub-family C (CFTR/MRP), member 8), CYP2C8(cytochrome P450, family 2, subfamily C, polypeptide 8), CYP3A4(cytochrome P450, family3, subfamily A, polypeptide 4), KCNJ11(potassium inwardly-rectifying channel, subfamily J, member 11), SLC01B1(solute carrier organic anion transporter family, member 1B1)	A10BX02
Riluzole	GRIN3A(glutamate receptor, ionotropic, N-methyl-D-aspartate 3)	CYP1A2(cytochrome P450, family 1, subfamily A, polypeptide 2), SCN5A(sodium channel, voltage-gated, type V, alpha subunit)	N07XX02
Cefdinir		MPO(myeloperoxidase)	J01DD15
Amcinonide	ANXA1(annexin A1),	NR3C1(nuclear receptor subfamily 3, group C, member 1)	D07AC11
Avandamet	POU6F1 (POU class 6 homeobox 1)		

Chapter II

Identification of Upstream Regulators for Prognostic Expression Signature Genes in Colorectal Cancer

ABSTRACT

Gene expression signatures have been commonly used as diagnostic and prognostic marker for cancer subtyping. However, expression signatures frequently include many *passengers*, which are not directly related to cancer progression. Their upstream regulators such as transcription factors (TFs) may take a more critical role as *driver* or master regulator, providing even better clues on underlying regulatory mechanism and therapeutic application. In order to indentify prognostic master regulators, we took the known 85 prognostic signature genes for colorectal cancer and inferred their upstream TFs. To this end, a global transcriptional regulatory network is constructed with total >200,000 TF-target links using ARACNE algorithm. We selected top 10 TFs as candidate master regulators, showing the highest coverage of the signature genes among the total 846 TF-target sub-networks or regulons. The selected TFs show a better prognostic performance than the original 85 signatures in spite of greatly reduced number of marker genes from 85 to 10. Notably, these TFs were selected solely on the basis of inferred regulatory links using gene expression profiles and include many TFs regulating tumorigenic processes such as proliferation, metastasis and differentiation. Our network approach lead to identification of the upstream transcription factors for prognostic signature genes, providing leads to their regulatory mechanism. We demonstrate that these TFs serve as even better biomarker by themselves than the original signature with markedly smaller size and better performance. The utility of our method may be expandable to other types of signatures such as diagnosis and drug response.

Keywords

Gene signature, colorectal cancer, transcriptional network, network inference

INTRODUCTION

With advances in genome-wide gene expression technologies, classification of cancer subtypes based on expression signatures has been widespread, resulting in plenty of biomarkers for various cancers. This molecular signature-based approach is generally more objective and reproducible than conventional methods based on clinico-pathological features. There are plenty of clinical applications being actively sought (1–3), some of which are already in commercial use (4, 5) for selecting treatment strategy and predicting prognosis. In spite of all those advantages and successful applications, identification of causal oncogenic pathways and driver regulators still remains as challenge (6). The main bottleneck is that expression signatures normally consist of both cancer drivers and passengers, where the latter is not directly related to cancer progression. Because passengers frequently take the majority of the signature genes, accurate discrimination of cancer drivers from passengers becomes a key subject of cancer genomic studies.

Regulatory network modeling has been widely used for systematic understanding of disease progression at the molecular level, particularly for cancer (comprehensively reviewed by Peer and Hachohen (7)). Recently, Carro and colleagues applied a reverse engineering method for context-specific transcriptional regulatory networks to 176 gene expression profiles from high grade glioblastoma (HGG) patients. Two TFs (C/EBP β and STAT3) were successfully identified as master regulators, controlling 'mesenchymal' signature genes leading to tumor aggressiveness such as epithelial-to-

mesenchymal transition and neo-angiogenesis (8). They applied ARACNE algorithm for global reconstruction of regulatory network (9), where directed or causal TF-target relationship is extracted by measuring conditional mutual information. Then, the regulatory TFs for the mesenchymal signature genes are inferred using master regulator analysis(MRA) together with or without stepwise linear regression method(SLR). It provides an exemplary case to pinpoint upstream regulators of known cancer signatures as cancer drivers and accordingly, promising therapeutic targets. Further, this strategy also provides a chance to develop biomarkers of even smaller size than the original signatures, which is highly desirable for practical usage in terms of both cost and interpretation.

In this work, we take the work by Carro and colleagues (8) as the framework of our analysis and apply essentially the same method to colorectal cancer with only minor modifications. Colorectal cancer is one of the most commonly diagnosed cancers and the fourth leading cause of cancer-related death in males and the third in females worldwide (10). Several research groups have identified prognostic molecular signatures using genome-wide gene expression profiles of colorectal cancer patients (11–13). Recently, Oh and colleagues reported a prognostic signature gene set consisting of 85 genes (114 probe sets) derived from the expression profiles of 177 colorectal cancer patients (14). This signature was able to discriminate colorectal cancer patients between good and poor prognostic groups with high accuracy. We reason that the upstream regulators or transcription factors (TFs) of these prognostic signatures may take a critical role as *driver* or master regulator, providing clues on the underlying regulatory mechanism and therapeutic

application. Here, we applied a reverse engineering algorithm to reconstruct an unbiased transcriptional network in colorectal cancer. Using this network, the upstream regulators of the prognostic signatures were identified and tested for their utility as prognostic marker. Our network models provided clues on the potential regulatory mechanisms how these upstream regulators cause prognostic differences.

RESULTS AND DISCUSSIONS

Overview of the analytic procedure

Our analytic procedure essentially follows that of Carro and colleagues for high-grade glioblastoma (HGG) (8), where a global regulatory network is inferred and the regulons or targets for each TF are extracted. The main difference is that the regulatory network modeling procedure is applied for the 85 prognostic marker genes for colorectal cancer, originally reported by Oh and colleagues (14) (Table II-1). Once the regulatory network is constructed, the top-ranked TFs are selected which regulate the highest number of signature genes. As the details of mathematical formulation is described in the previous work (9) and in Methods, we briefly summarize the overall procedure of our work here (Figure II-1). Once a global regulatory network is constructed using ARACNE algorithm, regulons or TF targets are extracted for all candidate TFs. Top candidate TFs are chosen on the basis of the coverage of signatures as downstream regulated genes (=regulons). This procedure or *master regulator analysis* (MRA) is essentially equal to conventional gene set analysis (GSA) based on Fisher's exact test. Alternatively, we applied a stepwise linear regression method (SLR) for each signature gene, where its expression is modeled using a minimal set of candidate TFs by linear regression. In our case, SLR was used only to filter out weak TF-target relations in each regulon, keeping only the most obvious interactions modeled by simple linear equation. In contrast, Carro and colleagues expanded the candidate TFs before applying SLR by including additional 52 TFs with their promoter sequence enriched among the signature

genes (8). Therefore, our work is more suitable to evaluate whether a regulatory model can successfully identify key upstream regulators (e.g. prognostic markers) purely based on expression profiles without depending on external knowledge such as TF binding sites.

Construction of regulatory networks and identification of upstream regulators for prognostic signatures

First, we took the 177 expression profiles from colon cancer patients from Moffit Cancer Center (Moffit cohort, $n=177$ (12)), which was also used to extract the prognostic 85 gene signature for colorectal cancer. Then ARACNE algorithm was applied to infer a global transcriptional network. In total, we inferred 155,818 TF-target interactions between 834 TFs and 17,065 target genes in the context of colorectal cancer (Figure II-1A). As a single TF and its targets constitute a *regulon* structure, the equal number of 834 regulons is extracted from the network (Figure II-1B). For the 834 regulons, we applied master regulator analysis (MRA), which tests significant overlap between the regulons and the 85 signature genes (Figure II-1C). MRA identified 67 TFs, of which targets significantly overlap with the signature at false discovery rate (FDR) < 0.05 (Table II-2). The 67 TFs collectively regulate 84 of the 85 signature genes, which are ranked by the order of signature coverage, i.e. the number of signature genes regulated by the corresponding TF. We took the top 10 TFs by MRA as candidates of master regulators for downstream analysis, which covers 83 of the 85 signature genes with the average number of targets per TF = 8.3 (Figure II-1D).

We further applied stepwise linear regression (SLR) to the ARACNE-generated network. In this step, the expression level of each signature gene was modeled by the linear combination of the expression levels of its upstream TFs in the network. Because SLR method tries to minimize the number of TFs in modeling the expression level of each signature gene, only the TFs showing strong linear correlation tend to be kept in the final regression model. Accordingly, SLR is essentially used as a filtering step to remove less effective TF-target interactions here. Then, the TFs were re-ranked by signature coverage using the remaining interactions after this filtering step. Again, we took the top 10 TFs by MRA+SLR, covering 71 of the 85 signature genes with the average number of targets per TF = 7.1 (Figure II-1E). The two TF sets between MRA and MRA+SLR largely agree to each other with 7 TFs in common (i.e. PLAGL2, PRRX1, SPDEF, SATB2, ASCL2, HIF1A, and TCF7). Three TFs were specific for MRA (BCL6, TFCP2L1, and FOSL2) and MRA+SLR (AEBP1, GTF2IRD1, and TCEAL1) respectively (Table II-4). These two sets of top 10 TFs by MRA and MRA+SLR method are finally chosen as candidate master regulators and named as TF_{MRA} and $TF_{MRA+SLR}$ respectively. As a result, two versions of regulatory networks are constructed between the top 10 upstream regulators (TF_{MRA} and $TF_{MRA+SLR}$) and their downstream targets among the 85 signature genes. The downstream signature genes of each TF are listed in Table II-5 and both networks for TF_{MRA} and $TF_{MRA+SLR}$ are visualized in Figure III-2. Notably, some transcription factors are linked by positive or negative regulatory interactions. ASCL2 is positively regulated by two TFs (PLAGL2 and TCF7) and negatively by SPDEF, suggesting a higher order structure among the

upstream regulators. Many of the prognostic signature genes are co-regulated by several TFs, e.g. ACSL6 by four TFs (TCF7, TCEAL1, SATB2 and HIF1A) and VAV3 by three TFs (GTF2IRD1, SATB2 and TCF7).

Prognostic effect analyses for the upstream regulators identified by MRA and MRA+SLR

The 85 signature genes consist of 34 low-risk and 51 high-risk marker genes, which are significantly up and down-regulated respectively in the patient group of better survival (14). Accordingly, we assigned the prognostic effect of the 67 TFs as positive (+) or negative (-) class depending on whether the majority of the downstream target genes are regulated in favor of expressing low-risk or high-risk signatures. First, we calculated Spearman's rank correlation between each TF and its downstream signature genes. The regulatory mode is determined by the sign of Spearman's rank correlation between a TF and its target, where positive correlation indicates 'activation' and negative does 'repression'. The prognostic effect of a TF is assigned positive (+) if the sum of activated low-risk and repressed high-risk genes is more than half among its downstream signature genes and vice versa. Among the 67 TFs selected by MRA, the prognostic effect of the 30 TFs was positive with the remaining 37 TFs being negative (Table II-3).

We focus on the top 10 TFs in TF_{MRA} and $TF_{MRA+SLR}$ and ask whether their prognostic effect is consistently observed across different data sources. Besides the Moffit cohort used for network construction by ARACNE, we took another set of gene expression profiles from Royal Melbourne Hospital (Melbourne cohort, n=95) (11). Positive prognostic effect was observed in

five out of the 10 TFs in TF_{MRA} and four in $TF_{MRA+SLR}$ in the Moffit cohort (Figure II-3A). The rest five and six TFs show negative prognostic effect respectively. We observe exactly the same trend for all the TFs tested in the Melbourne cohort, suggesting that their regulatory interactions are consistently maintained in colorectal cancer (Figure II-3B).

Strong association of the top 10 upstream TFs with the survival of colon cancer patients

Now, we test the utility of the upstream regulators (TF_{MRA} , $TF_{MRA+SLR}$) as prognostic marker for colorectal cancer. In the Moffit cohort (n=177) used as training dataset, both TF_{MRA} and $TF_{MRA+SLR}$ show a strong differential expression pattern between the good and the poor prognostic group similar to that of the original 85 signature genes (Figure II-4A, II-4B). An SVM (support vector machine) classifier was constructed for TF_{MRA} , $TF_{MRA+SLR}$, and the original 85 signature genes. For validation purpose, we took the Melbourne cohort (n=95) as an independent test set. These 95 patients were classified into *good* or *poor* prognostic groups independently using each of the three classifiers. For all three classifications, the resulting good and poor prognostic groups show the same differential expression patterns in the test dataset as well (Figure II-4C, II-4D and II-4E).

We compared the prognostic performance of the three classifiers using the Kaplan-Meier plots for disease-free survival (Figure II-5). The upstream TFs show a slightly better or similar performance than the original 85 gene signature with the ordering of $TF_{MRA+SLR} > TF_{MRA} > 85$ gene signature. The P-values by log-rank test were 1.97×10^{-3} for $TF_{MRA+SLR}$, 5.15×10^{-3} for

TF_{MRA} and 5.15×10^{-3} for the 85 gene signature. Notably, these upstream TFs were not selected directly for good (or poor) survival but by the coverage of known prognostic signatures in our regulatory network model based purely on expression profiles. Therefore, the performances of TF_{MRA} and $TF_{MRA+SLR}$ are thought to be unexpectedly high, considering that the signature size dramatically decreased to less than 1/8 (from 85 to 10 genes). It demonstrates that the upstream TFs can be even better prognostic marker than the expression signature itself. The same strategy may be useful in identifying upstream regulators for other types of cancer signatures such as drug response and metastatic behavior.

Candidate upstream regulators include many TFs involved in tumorigenesis: HIF1A FOSL2, PLAGL2, ASCL2, and TCF7

Many of the upstream TFs for the prognostic signature genes are actually well-known regulators for various tumorigenic processes such as cell invasion, metastasis and clinical outcomes. Among the TFs of poor prognostic effect, HIF1A and FOSL2 are such cases. Our network models also recapitulate some of the known TF-target relations, which are confirmed by literature. Hypoxia-inducible factors (HIFs) are the key regulators of oxygen signaling pathway responding to oxygen-deficient environment called as hypoxia. Cancer cells overcome hypoxic condition by hypoxic pathway activated by HIFs. HIF1A is overexpressed in a variety of human cancers, which is associated with poor prognosis in various cancers (15, 16) including colon cancer (17). Among the nine targets of HIF1A in our network by MRA+SLR, three interactions are confirmed by literature. HIF1A activates

CXCR4 and LOX, which are involved in metastasis in renal cell carcinoma (18) and hypoxia-induced metastasis(19) respectively. PTGS2 (known as COX2) is known to be directly up-regulated by HIF1A and promotes hypoxia-induced angiogenesis (20). Also, PTGS2 is shown negatively regulated by ASCL2, one among the top 10 TFs in both networks. FOSL2 (also known as FRA2) is a member of FOS family, which encode leucine zipper proteins forming AP-1 transcription factor complex together with JUN family proteins. While FOSL2 is included in the top 10 TFs only in TF_{MRA} , its rank is still relatively high in $TF_{MRA+SLR}$ (19th out of the 67 TFs). FOSL2 is known to mediate cell growth and differentiation (21) and its transgenic mice show a severe loss of small blood vessels in skin (22), suggesting a role in angiogenesis. FOSL2 also activates LOX in our network by MRA (Figure II-1A).

Among the TFs of good prognostic effect, PLAGL2 is notable due to its dual functionalities as proto-oncogene and tumor suppressor. PLAGL2 has been known as a proto-oncogene in acute myeloid leukemia (AML), glioblastoma (GBM) and colorectal cancer (23–25). PLAGL2 can activate Wnt signaling, leading to leukemia in mice (23) or suppression of cellular differentiation (25). On the contrary, PLAGL2 also function as tumor suppressor by promoting apoptosis or arresting cell cycle (26–28). ASCL2 and TCF7 (also known as TCF-1) are TFs activated by Wnt signaling. ASCL2 is up-regulated in colorectal adenocarcinoma (29) and recently, growth arrest is observed by knockdown of ASCL2 in vivo (30) although the prognostic effect of ASCL2 is positive(+). TCF7 is a member of the TCF/LEF family, which transmit Wnt signal into the nucleus and activate Wnt target genes by

interacting with β -catenin. Unlike other members of TCF/LEF family, TCF7 may act as negative regulators for Wnt signaling because its isoforms lack β -catenin binding domain, while retaining Groucho interaction domain necessary for repressor activity (31, 32).

There are evidences for tumorigenic activity for other TFs such as PRRX1 (PMX1) and SPDEF (PDEF). The gene fusion between PRRX1 and NUP98 was reported in AML (33). Suppressive activities for metastasis, cell growth and migration are suggested for SPDEF (34, 35).

CONCLUSIONS

We propose a genetic analysis pipeline to find transcriptional modules for prognostic gene expression signatures or other biomarkers. Our method only requires expression profiles in the appropriate context such as tissue type and disease condition. This procedure is applied to identify key upstream regulators for the 85 prognostic signature genes for colorectal cancer. To this end, a global transcriptional network is constructed using ARACNE algorithm (9). Candidate upstream regulators are selected on the basis of the number of signature genes as downstream targets or regulons (MRA step). Additional filter was applied to extract only strong TF-target interactions readily modeled by simple linear regression (SLR step). As a result, we identified two sets of top 10 TFs, which clearly discriminate between the good and poor prognostic group. The prognostic performance is tested using a dataset independent of both signature selection and network modeling. These candidate upstream TFs include many known regulators for tumorigenic processes such as metastasis and cell proliferation. The utility of our work is two-fold. One is to allow the identification of upstream regulators for a given set of signature genes, providing leads to regulatory mechanism. Second, these regulators may serve as even better biomarker by themselves than the original signatures with markedly smaller size and better performance. The utility of our method may be expandable to other types of signatures such as diagnosis and drug response.

METHODS

Data set

The 85 prognostic gene signature for colorectal cancer is obtained from S-C Oh et al., which was derived by mapping the 114 probes to the corresponding genes (14). The gene expression profiles from Moffit cohort (GSE17536, n=177) and those from Melbourne cohort (GSE14333, n=95 after removal of redundancy) were obtained from Gene Expression Omnibus database (<http://ncbi.nlm.nih.gov/geo>). All the expression profiles used were generated using Affymetrix HG-U133 Plus2.0 GeneChip array. The raw CEL files were processed and normalized using the MAS5 method (affy package in R/Bioconductor). The list of TFs was obtained from Carro et al.(8), which includes 928 human TFs. These TFs were mapped to 2155 probe sets in Affymetrix HG-U133Plus2.0 GeneChip array.

Network inference using ARACNE

ARACNE

(<http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE>)

was used to infer interactions between the 2155 TF probe sets and their target genes. The gene expression profiles of the Moffit cohort were used in this analysis. Threshold for MI (mutual information) and DPI (Data Processing Inequality) tolerance were set to $p < 0.05$ (Bonferroni corrected for multiple testing) and 0%, respectively. Bootstrapping option was applied to generate 100 bootstrap networks. These networks were merged into a consensus network by consensus voting method based on statistically significant number

of interactions inferred in bootstrapping steps. As probe sets in network were mapped to genes, the consensus network was merged into gene level network.

Master regulator analysis

Fisher's exact test was used in order to determine statistical significance for overlaps between target genes in each regulon. The FDRs for the p-values are computed using procedures described by Benjamini and Hochberg (36). Then, the signature-enriched TFs were ranked by signature coverage, which is the edge number linked with signature genes.

Stepwise linear regression analysis

A linear model for each signature gene was constructed as follows. The log2-expression level of TFs linked to each signature gene was considered as the explanatory variables. The log2-expression level of each signature gene was considered as the response variable. Then we used stepwise algorithm in order to select the best minimal set of the explanatory variables in each model. Akaike information criterion (AIC) was used as stop criterion. TFs that the p-value for linear regression coefficient was less than 0.05 were removed in selected variables.

Class prediction and survival analysis

BRB-Array Tools (<http://linus.nci.nih.gov/BRB-ArrayTools.html>) was used for building SVM classifier and class prediction. The survival package in R was used for Kaplan-Meier plot and log-rank test.

ABBREVIATIONS

MRA: master regulator analysis

SLR: stepwise linear regression

AML: acute myeloid leukemia;

GBM: glioblastoma;

TCF: T-cell factor;

LEF: lymphoid enhancer factor

REFERENCES

1. Sotiriou,C. and Piccart,M.J. (2007) Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nature reviews. Cancer*, **7**, 545–53.
2. Méndez,E., Lohavanichbutr,P., Fan,W., Houck,J.R., Rue,T.C., Doody,D.R., Futran,N.D., Upton,M.P., Yueh,B., Zhao,L.P., et al. (2011) Can a metastatic gene expression profile outperform tumor size as a predictor of occult lymph node metastasis in oral cancer patients? *Clinical Cancer Research*, **17**, 2466–2473.
3. Servant,N., Bollet,M.A., Halfwerk,H., Bleakley,K., Kreike,B., Jacob,L., Sie,D., Kerkhoven,R., Hupe,P., Hadhri,R., et al. (2012) Search for a gene expression signature of breast cancer local recurrence in young women. *Clinical Cancer Research*, **45**, 1704–15.
4. Veer,L.J. Van, Dai,H., Vijver,M.J. Van De, Schreiber,G.J., Kerkhoven,R.M., Roberts,C., Bernards,Â., Friend,S.H. and Linsley,P.S. (2002) Gene expression profiling predicts clinical outcome of breast cancer. **415**.
5. Paik,S., Shak,S., Tang,G., Kim,C., Baker,J., Cronin,M., Baehner,F.L., Walker,M.G., Watson,D., Park,T., et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.
6. Nevins,J.R. and Potti,A. (2007) Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature reviews. Genetics*, **8**, 601–9.
7. Pe’er,D. and Hacohen,N. (2011) Principles and strategies for

- developing network models in cancer. *Cell*, **144**, 864–73.
8. Carro,M.S., Lim,W.K., Alvarez,M.J., Bollo,R.J., Zhao,X., Snyder,E.Y., Sulman,E.P., Anne,S.L., Doetsch,F., Colman,H., et al. (2010) The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, **463**, 318–25.
 9. Margolin,A. a, Wang,K., Lim,W.K., Kustagi,M., Nemenman,I. and Califano,A. (2006) Reverse engineering cellular networks. *Nature protocols*, **1**, 662–71.
 10. Jemal,A., Bray,F., Center,M.M., Ferlay,J., Ward,E. and Forman,D. (2011) Global cancer statistics. *CA: a cancer journal for clinicians*, **61**, 69–90.
 11. Jorissen,R.N., Gibbs,P., Christie,M., Prakash,S., Lipton,L., Desai,J., Kerr,D., Aaltonen,L.A., Arango,D., Kruhøffer,M., et al. (2009) Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. *Clinical cancer research: an official journal of the American Association for Cancer Research*, **15**, 7642–7651.
 12. Smith,J.J., Deane,N.G., Wu,F., Merchant,N.B., Zhang,B., Jiang,A., Lu,P., Johnson,J.C., Schmidt,C., Bailey,C.E., et al. (2010) Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*, **138**, 958–68.
 13. Staub,E., Groene,J., Heinze,M., Mennerich,D., Roepcke,S., Klamann,I., Hinzmann,B., Castanos-Velez,E., Pilarsky,C., Mann,B., et al. (2009) An expression module of WIPF1-coexpressed genes identifies

- patients with favorable prognosis in three tumor types. *Journal of molecular medicine (Berlin, Germany)*, **87**, 633–44.
14. Oh,S.C., Park,Y.-Y., Park,E.S., Lim,J.Y., Kim,S.M., Kim,S.-B., Kim,J., Kim,S.C., Chu,I.-S., Smith,J.J., et al. (2012) Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. *Gut*, **61**, 1291–8.
 15. Rankin,E.B. and Giaccia, a J. (2008) The role of hypoxia-inducible factors in tumorigenesis. *Cell death and differentiation*, **15**, 678–85.
 16. Majmundar,A.J., Wong,W.J. and Simon,M.C. (2010) Hypoxia-inducible factors and the response to hypoxic stress. *Molecular cell*, **40**, 294–309.
 17. Baba,Y., Noshok,K., Shima,K., Irahara,N., Chan,A.T., Meyerhardt,J. a., Chung,D.C., Giovannucci,E.L., Fuchs,C.S. and Ogino,S. (2010) HIF1A Overexpression Is Associated with Poor Prognosis in a Cohort of 731 Colorectal Cancers. *The American Journal of Pathology*, **176**, 2292–2301.
 18. Pan,J., Mestas,J., Burdick,M.D., Phillips,R.J., Thomas,G. V, Reckamp,K., Belperio,J. a and Strieter,R.M. (2006) Stromal derived factor-1 (SDF-1/CXCL12) and CXCR4 in renal cell carcinoma metastasis. *Molecular cancer*, **5**, 56.
 19. Erler,J.T., Bennewith,K.L., Nicolau,M., Dornhöfer,N., Kong,C., Le,Q.-T., Chi,J.-T.A., Jeffrey,S.S. and Giaccia,A.J. (2006) Lysyl oxidase is essential for hypoxia-induced metastasis. *Nature*, **440**, 1222–6.
 20. Zhong,H., Willard,M. and Simons,J. (2004) NS398 reduces hypoxia-

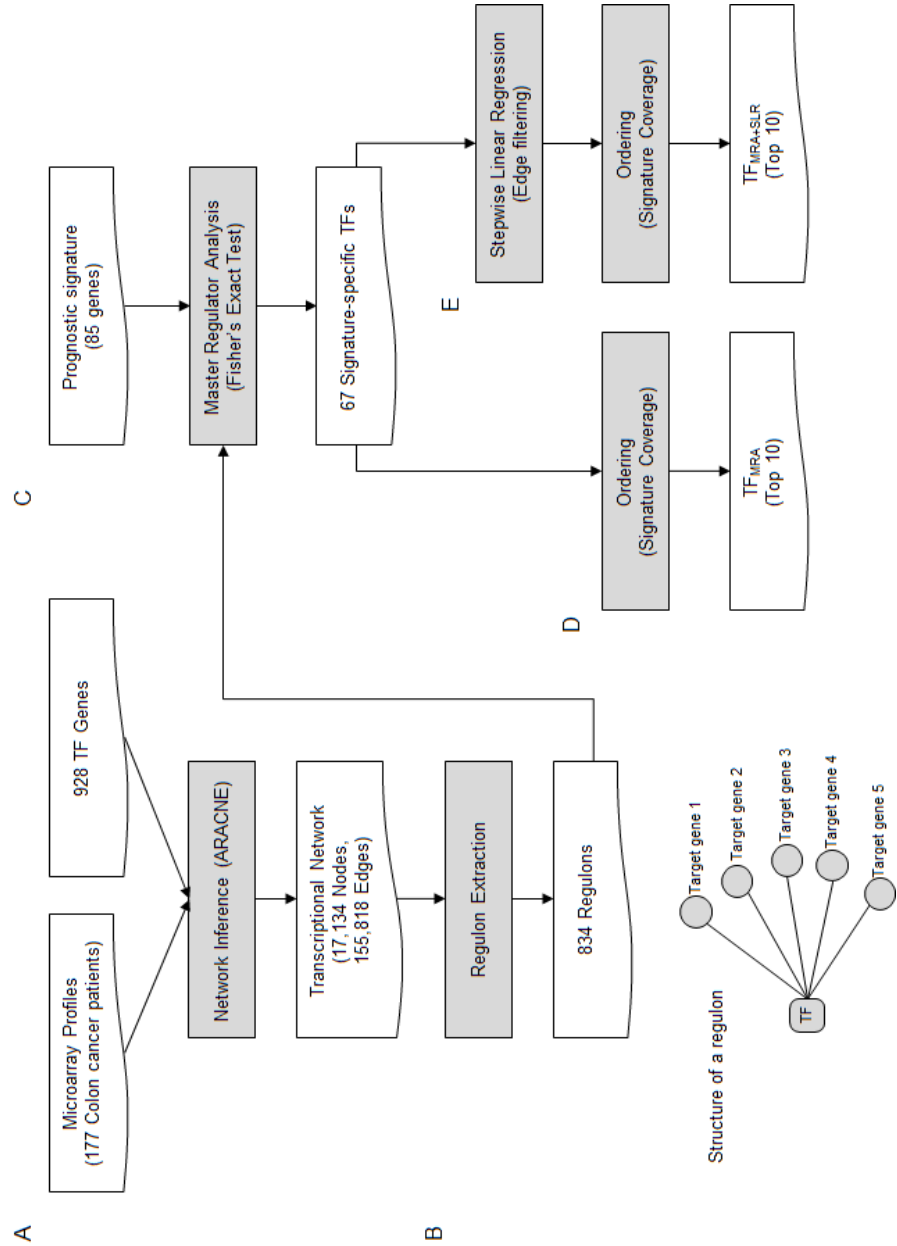
- inducible factor (HIF)-1 α and HIF-1 activity: multiple-level effects involving cyclooxygenase-2 dependent and independent mechanisms. *International journal of cancer. Journal international du cancer*, **112**, 585–95.
21. Outinen,P. a, Sood,S.K., Pfeifer,S.I., Pamidi,S., Podor,T.J., Li,J., Weitz,J.I. and Austin,R.C. (1999) Homocysteine-induced endoplasmic reticulum stress and growth arrest leads to specific changes in gene expression in human vascular endothelial cells. *Blood*, **94**, 959–67.
 22. Maurer,B., Busch,N., Jüngel,A., Pileckyte,M., Gay,R.E., Michel,B. a, Schett,G., Gay,S., Distler,J. and Distler,O. (2009) Transcription factor fos-related antigen-2 induces progressive peripheral vasculopathy in mice closely resembling human systemic sclerosis. *Circulation*, **120**, 2367–76.
 23. Landrette,S.F., Kuo,Y.-H., Hensen,K., Barjesteh van Waalwijk van Doorn-Khosrovani,S., Perrat,P.N., Van de Ven,W.J.M., Delwel,R. and Castilla,L.H. (2005) Plag1 and Plag12 are oncogenes that induce acute myeloid leukemia in cooperation with Cbfb-MYH11. *Blood*, **105**, 2900–7.
 24. Yang,Y.-S., Yang,M.-C.W. and Weissler,J.C. (2011) Pleiomorphic adenoma gene-like 2 expression is associated with the development of lung adenocarcinoma and emphysema. *Lung cancer (Amsterdam, Netherlands)*, **74**, 12–24.
 25. Zheng,H., Ying,H., Wiedemeyer,R., Yan,H., Quayle,S.N., Ivanova,E. V, Paik,J.-H., Zhang,H., Xiao,Y., Perry,S.R., et al. (2010) PLAGL2 regulates Wnt signaling to impede differentiation in neural stem cells

- and gliomas. *Cancer cell*, **17**, 497–509.
26. Furukawa,T., Adachi,Y., Fujisawa,J., Kambe,T., Yamaguchi-Iwai,Y., Sasaki,R., Kuwahara,J., Ikehara,S., Tokunaga,R. and Taketani,S. (2001) Involvement of PLAGL2 in activation of iron deficient- and hypoxia-induced gene expression in mouse cell lines. *Oncogene*, **20**, 4718–27.
 27. Mizutani,A., Furukawa,T., Adachi,Y., Ikehara,S. and Taketani,S. (2002) A zinc-finger protein, PLAGL2, induces the expression of a proapoptotic protein Nip3, leading to cellular apoptosis. *The Journal of biological chemistry*, **277**, 15851–8.
 28. Hanks,T.S. and Gauss,K. a (2012) Pleomorphic adenoma gene-like 2 regulates expression of the p53 family member, p73, and induces cell cycle block and apoptosis in human promonocytic U937 cells. *Apoptosis : an international journal on programmed cell death*, **17**, 236–47.
 29. Jubb, a M., Chalasani,S., Frantz,G.D., Smits,R., Grabsch,H.I., Kavi,V., Maughan,N.J., Hillan,K.J., Quirke,P. and Koeppen,H. (2006) Achaete-scute like 2 (ascl2) is a target of Wnt signalling and is upregulated in intestinal neoplasia. *Oncogene*, **25**, 3445–57.
 30. Zhu,R., Yang,Y., Tian,Y., Bai,J., Zhang,X., Li,X., Peng,Z., He,Y., Chen,L., Pan,Q., et al. (2012) Ascl2 Knockdown Results in Tumor Growth Arrest by miRNA-302b-Related Inhibition of Colon Cancer Progenitor Cells. *PloS one*, **7**, e32170.
 31. Roose,J. (1999) Synergy Between Tumor Suppressor APC and the - Catenin-Tcf4 Target Tcf1. *Science*, **285**, 1923–1926.

32. Waterman,M.L. (2004) Lymphoid enhancer factor/T cell factor expression in colorectal cancer. *Cancer and Metastasis Reviews*, **23**, 41–52.
33. Nakamura,T., Yamazaki,Y., Hatano,Y. and Miura,I. (1999) NUP98 is fused to PMX1 homeobox gene in human acute myelogenous leukemia with chromosome translocation t(1;11)(q23;p15). *Blood*, **94**, 741–7.
34. Moussa,O., Turner,D.P., Feldman,R.J., Sementchenko,V.I., McCarragher,B.D., Desouki,M.M., Fraig,M. and Watson,D.K. (2009) PDEF is a negative regulator of colon cancer cell growth and migration. *Journal of cellular biochemistry*, **108**, 1389–98.
35. Steffan,J.J. and Koul,H.K. (2011) Prostate derived ETS factor (PDEF): a putative tumor metastasis suppressor. *Cancer letters*, **310**, 109–17.
36. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, **57**, 289–300.

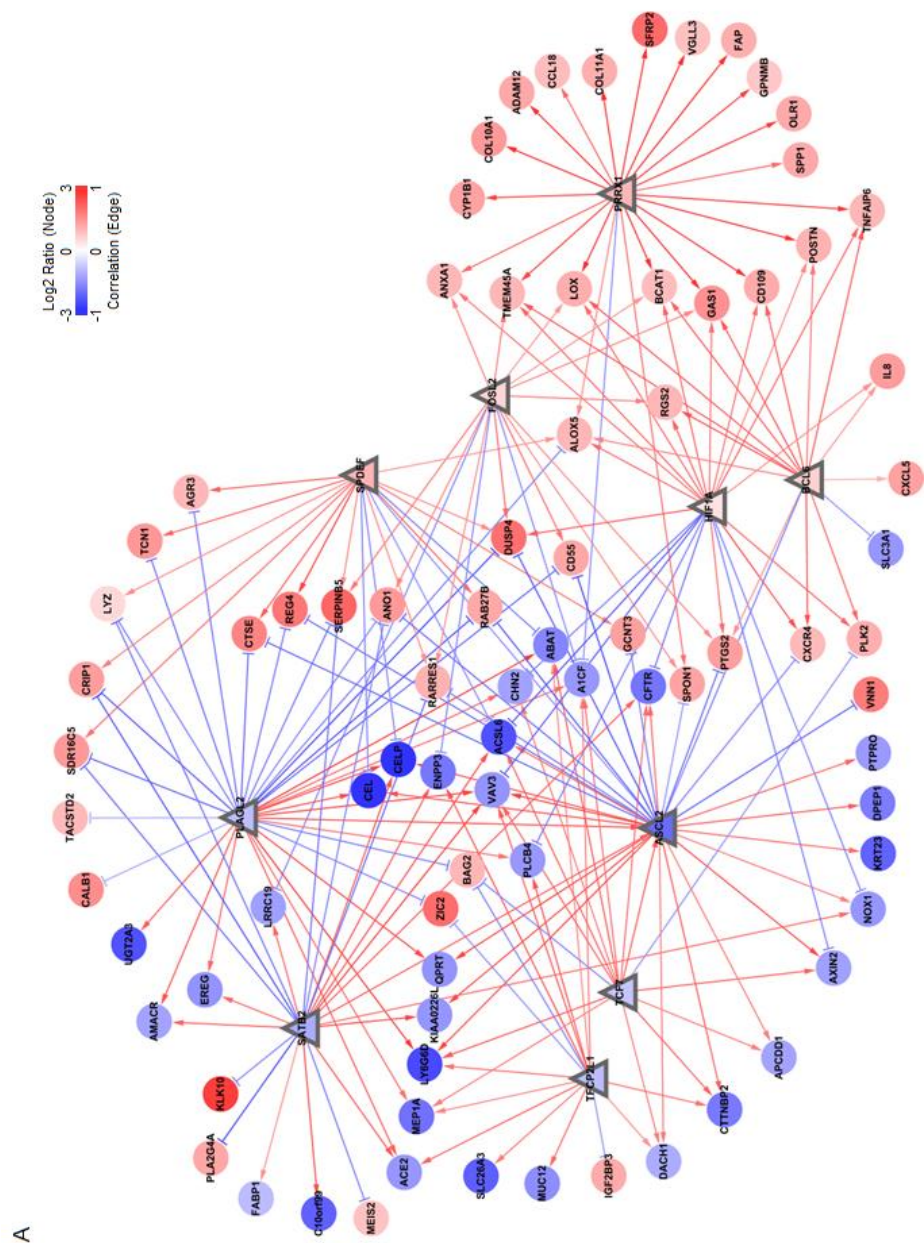
FIGURES AND TABLES

Figure II-1. Overall pipeline of upstream regulator inference



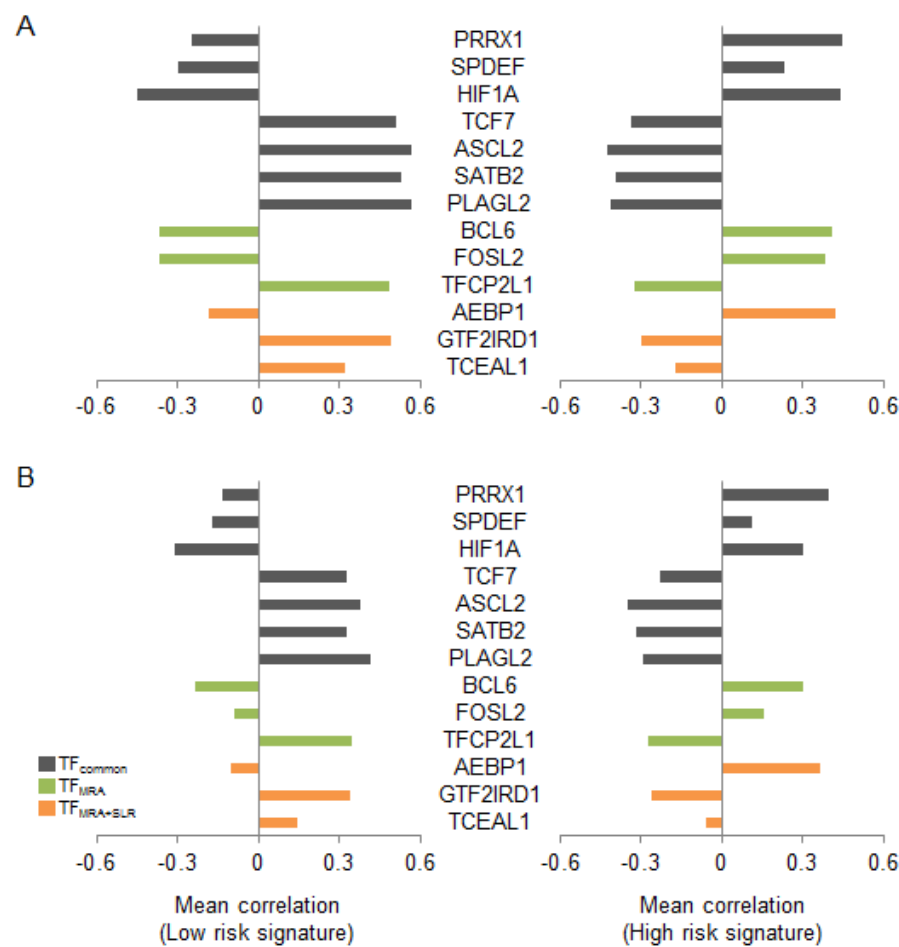
(A) Global regulatory network modeling using ARACNE. (B) Regulon extraction for each TF. (C) Master regulator analysis (MRA) selects the TFs showing a significant overlap with the prognostic signature genes (D) Extraction of top 10 TFs by the signature coverage of MRA derived regulons (E) Stepwise linear regression (SLR) for edge filtering and extraction of top 10 TFs by the signature coverage of MRS+SLR derived regulons.

Figure II-2. The transcriptional network between the top 10 TFs and the signature genes



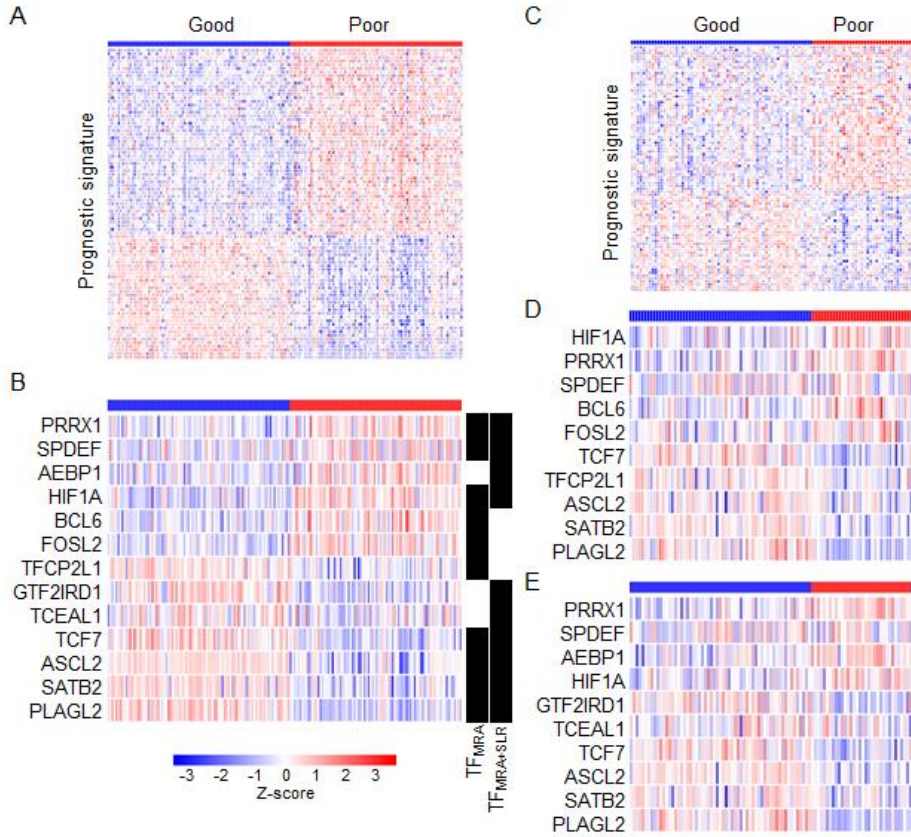
cohort (n=177). Arrow shape of edge represents regulatory mode determined by the sign(+/-) of Spearman's rank correlation between a TF and its target gene. Edge color represents the magnitude of correlation.

Figure II-3. Correlation between the upstream TFs and their target genes



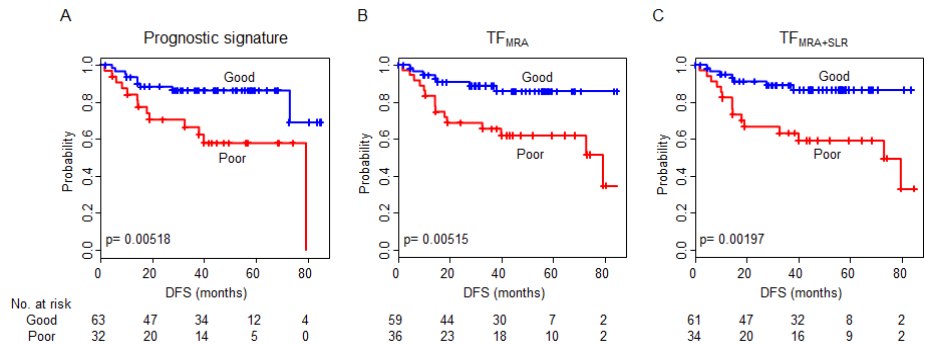
Mean Spearman's rank correlations are calculated between the 13 TFs (union of TF_{MRA} and TF_{MRA+SLR}) and the low risk (left) or the high risk (right) signature genes for (A) Moffit cohort and (B) Melbourne cohort.

Figure II-4. Expression patterns of the selected marker genes between the good and the poor prognostic group



The distinct expression pattern of (A) 85 signature genes and of (B) 13 TFs (union of TF_{MRA} and $TF_{MRA+SLR}$) are shown in the Moffit cohort (n=177, training dataset). Differential expression pattern is observed to be well maintained in an independent test dataset (Melbourne cohort, n=95) for (C) the 85 signature genes, (D) TF_{MRA} and (E) $TF_{MRA+SLR}$ after class prediction.

Figure II-5. The prediction performance of the selected prognostic markers



Kaplan-Meier plots for disease-free survival (DFS) are shown between the good and the poor prognostic group for (A) the 85 signature genes, (B) TF_{MRA} , and (C) $TF_{MRA+SLR}$. P value for difference between two K-M plots was calculated by log-rank test.

Table II-1. 85 Signature genes

TF symbol
A1CF, ABAT, ACE2, ACSL6, ADAM12, AGR3, AHNAK2, ALOX5, AMACR, ANO1, ANXA1, APCDD1, ASCL2, AXIN2, BAG2, BCAT1, C10orf99, CALB1, CCL18, CD109, CD55, CEL, CELP, CFTR, CHN2, COL10A1, COL11A1, CRIP1, CTSE, CTTNBP2, CXCL5, CXCR4, CYP1B1, DACH1, DEFA6, DPEP1, DUSP4, ENPP3, EREG, FABP1, FAP, GAS1, GCNT3, GPNMB, IGF2BP3, IL8, KIAA0226L, KLK10, KRT23, LOX, LRRC19, LY6G6D, LYZ, MEIS2, MEP1A, MUC12, NOX1, OLR1, PLA2G4A, PLCB4, PLK2, POSTN, PTGS2, PTPRO, QPRT, RAB27B, RARRES1, REG4, RGS2, SDR16C5, SERPINB5, SFRP2, SLC26A3, SLC3A1, SPON1, SPP1, TACSTD2, TCN1, TMEM45A, TNFAIP6, UGT2A3, VAV3, VGLL3, VNN1, ZIC2

Table II-2. Master Regulator Candidates

TF symbol	Prognostic effect	Regulon size	MRA			MRA+SLR	
			Rank	Signature coverage	FDR adjusted p value	Rank	Signature coverage
PLAGL2	+	575	1	32	7.73E-25	1	20
PRRX1	-	327	4	22	2.10E-18	2	17
SPDEF	-	304	6	18	6.50E-14	3	16
SATB2	+	264	3	23	1.28E-21	4	15
ASCL2	+	537	2	28	6.12E-21	5	11
AEBP1	-	465	15	12	2.21E-05	6	10
HIF1A	-	371	4	22	2.58E-17	7	9
TCF7	+	408	9	15	9.47E-09	7	9
GTF2IRD1	+	429	16	10	0.000334	7	9
TCEAL1	+	276	16	10	1.08E-05	7	9
BCL6	-	455	7	16	4.50E-09	11	8
CDX2	+	233	12	13	1.52E-09	11	8
CBFA2T2	+	819	16	10	0.0308	11	8
TFCP2L1	+	364	9	15	2.26E-09	14	7
ETS2	+	363	16	10	8.50E-05	14	7
ZKSCAN1	+	497	16	10	0.000935	14	7
SLC26A3	+	187	24	9	4.86E-06	14	7
CREB3L1	-	197	29	7	0.000444	14	7
FOSL2	-	421	7	16	1.70E-09	19	6
ELK3	-	212	12	13	5.18E-10	19	6
NR1I2	+	204	16	10	8.60E-07	19	6
ZNF91	+	249	11	14	2.73E-10	22	5
ARNTL2	-	272	16	10	9.98E-06	22	5
CREM	-	255	25	8	0.000325	22	5
PLAGL1	+	109	29	7	1.60E-05	22	5
SCML1	+	252	29	7	0.00186	22	5
ZNF43	+	143	29	7	7.54E-05	22	5
FOXD1	-	97	36	6	9.43E-05	22	5
HIVEP3	-	201	36	6	0.0039	22	5
HOXB9	+	59	36	6	8.62E-06	22	5
PPARD	-	88	36	6	6.40E-05	22	5
TCF7L2	-	265	36	6	0.0143	22	5
MAFB	-	221	46	5	0.0321	22	5
ETV5	-	146	29	7	8.23E-05	34	4
NFIB	+	292	29	7	0.00431	34	4
JUN	+	203	46	5	0.0249	34	4
MEOX2	-	157	46	5	0.00904	34	4
SALL1	+	77	46	5	0.000444	34	4
SNAI2	-	172	46	5	0.0131	34	4
TFDP2	+	218	46	5	0.0308	34	4

EGR2	-	68	54	4	0.00377	34	4
NR4A2	-	38	54	4	0.000444	34	4
ZNF193	+	119	54	4	0.0244	34	4
ZBTB38	+	426	12	13	1.42E-06	44	3
LMO4	-	306	25	8	0.000935	44	3
ZNF281	-	382	29	7	0.0177	44	3
AHR	-	88	36	6	6.40E-05	44	3
ETV1	-	131	36	6	0.000444	44	3
ZNF532	-	226	46	5	0.0348	44	3
ESRRG	+	37	54	4	0.00044	44	3
SCML2	+	42	54	4	0.000637	44	3
STAT2	-	333	16	10	5.18E-05	52	2
RUNX2	-	577	25	8	0.0381	52	2
ELF1	+	310	36	6	0.028	52	2
TFAP2A	-	156	36	6	0.00104	52	2
HHEX	-	78	54	4	0.00573	52	2
MEIS2	-	106	54	4	0.0165	52	2
ZNF267	-	72	62	3	0.0441	52	2
ZNF273	+	57	62	3	0.0257	52	2
INSM1	-	22	65	2	0.0477	52	2
ZNF185	-	10	65	2	0.0129	52	2
PHTF2	-	395	25	8	0.0047	62	1
SOX11	-	84	36	6	5.29E-05	62	1
NFE2L3	+	141	46	5	0.00573	62	1
HOXC6	-	56	62	3	0.0249	62	1
ISL1	-	23	54	4	7.54E-05	66	0
HOXD11	+	17	65	2	0.0308	66	0

Table II-3. Prognostic Effect Summary

TF	High risk signature		Low risk signature		Prognostic effect
	+	-	+	-	
PLAGL2	0	17	15	0	+
PRRX1	21	0	0	1	-
SPDEF	13	0	0	5	-
SATB2	0	9	14	0	+
ASCL2	0	12	16	0	+
AEBP1	12	0	0	0	-
HIF1A	15	0	0	7	-
TCF7	0	2	13	0	+
GTF2IRD1	0	1	9	0	+
TCEAL1	0	3	7	0	+
BCL6	14	0	0	2	-
CDX2	0	7	6	0	+
CBFA2T2	0	5	5	0	+
TFCP2L1	0	3	12	0	+
ETS2	0	2	8	0	+
ZKSCAN1	0	0	10	0	+
SLC26A3	0	0	9	0	+
CREB3L1	6	0	0	1	-
FOSL2	13	0	0	3	-
ELK3	12	0	0	1	-
NR1I2	0	2	8	0	+
ZNF91	0	10	4	0	+
ARNTL2	4	0	0	6	-
CREM	8	0	0	0	-
PLAGL1	0	3	4	0	+
SCML1	0	2	5	0	+
ZNF43	0	3	4	0	+
FOXD1	4	0	0	2	-
HIVEP3	6	0	0	0	-
HOXB9	0	1	5	0	+
PPARD	4	0	0	2	-
TCF7L2	6	0	0	0	-
MAFB	5	0	0	0	-
ETV5	3	0	0	4	-
NFIB	0	2	5	0	+
JUN	0	2	3	0	+
MEOX2	5	0	0	0	-
SALL1	0	3	2	0	+
SNAI2	5	0	0	0	-
TFDP2	0	3	2	0	+
EGR2	4	0	0	0	-
NR4A2	4	0	0	0	-

ZNF193	0	3	1	0	+
ZBTB38	0	4	9	0	+
LMO4	2	0	0	6	-
ZNF281	6	0	0	1	-
AHR	3	0	0	3	-
ETV1	6	0	0	0	-
ZNF532	5	0	0	0	-
ESRRG	0	0	4	0	+
SCML2	0	2	2	0	+
STAT2	3	0	0	7	-
RUNX2	8	0	0	0	-
ELF1	0	1	5	0	+
TFAP2A	3	0	0	3	-
HHEX	2	0	0	2	-
MEIS2	4	0	0	0	-
ZNF267	3	0	0	0	-
ZNF273	0	3	0	0	+
INSM1	2	0	0	0	-
ZNF185	2	0	0	0	-
PHTF2	7	0	0	1	-
SOX11	6	0	0	0	-
NFE2L3	0	2	3	0	+
HOXC6	1	0	0	2	-
ISL1	4	0	0	0	-
HOXD11	0	0	2	0	+

Table II-4. Overall statistics of 13 TFs, union of TF_{MRA} and TF_{MRA+SLR}

TF symbol	Prognostic effect	Regulon size	MRA ¹			MRA+SLR ²	
			Rank	Signature coverage	FDR ³	Rank	Signature coverage
PLAGL2	+	575	1	32	7.73E-25	1	20
PRRX1	-	327	4	22	2.10E-18	2	17
SPDEF	-	304	6	18	6.50E-14	3	16
SATB2	+	264	3	23	1.28E-21	4	15
ASCL2	+	537	2	28	6.12E-21	5	11
AEBP1	-	465	15	12	2.21E-05	6	10
GTF2IRD1	+	429	16	10	0.000334	7	9
HIF1A	-	371	4	22	2.58E-17	7	9
TCEAL1	+	276	16	10	1.08E-05	7	9
TCF7	+	408	9	15	9.47E-09	7	9
BCL6	-	455	7	16	4.50E-09	11	8
TFCP2L1	+	364	9	15	2.26E-09	14	7
FOSL2	-	421	7	16	1.70E-09	19	6

¹ Master Regulatory Analysis² Stepwise Linear Regression³ False Discovery Rate

Table II-5. Top 10 TF-Target list of MRA and MRA+SLR method

TF symbol	Target genes (MRA)	Target genes (MRA+SLR)
PRRX1	A1CF, ADAM12, ALOX5, ANXA1, BCAT1, CCL18, CD109, COL10A1, COL11A1, CYP1B1, FAP, GAS1, GPNMB, LOX, OLR1, POSTN, SFRP2, SPON1, SPP1, TMEM45A, TNFAIP6, VGLL3	ADAM12, BCAT1, CCL18, CD109, COL10A1, COL11A1, CYP1B1, FAP, GAS1, LOX, OLR1, POSTN, SFRP2, SPON1, SPP1, TNFAIP6, VGLL3
SPDEF	ABAT, ACSL6, AGR3, ALOX5, ASCL2, CEL, CELP, CRIP1, CTSE, DUSP4, GCNT3, LYZ, RAB27B, RARRES1, REG4, SDR16C5, SERPINB5, TCN1	ABAT, AGR3, ALOX5, ASCL2, CEL, CELP, CRIP1, CTSE, GCNT3, LYZ, RAB27B, RARRES1, REG4, SDR16C5, SERPINB5, TCN1
AEBP1		ADAM12, COL10A1, COL11A1, FAP, GAS1, GPNMB, MEIS2, SFRP2, SPON1, TMEM45A
HIF1A	ACSL6, ALOX5, ANXA1, AXIN2, BCAT1, CD109, CFTR, CHN2, CXCR4, DUSP4, GAS1, IL8, LOX, NOX1, PLCB4, PLK2, POSTN, PTGS2, RGS2, TMEM45A, TNFAIP6, VAV3	ACSL6, ALOX5, BCAT1, CXCR4, LOX, PLK2, PTGS2, RGS2, TMEM45A
BCL6	ALOX5, ASCL2, BCAT1, CD109, CXCL5, CXCR4, GAS1, IL8, LOX, PLK2, POSTN, PTGS2, RGS2, SLC3A1, TMEM45A, TNFAIP6	
FOSL2	A1CF, ANO1, ANXA1, BCAT1, CD55, DUSP4, ENPP3, GAS1, LOX, LRRC19, PTGS2, RARRES1, RGS2, SERPINB5, SPON1, TMEM45A	
TFCP2L1	A1CF, ABAT, ACE2, BAG2, CTTNBP2, DACH1, ENPP3, IGF2BP3, LY6G6D, MEP1A, MUC12, PLCB4, SLC26A3, VAV3, ZIC2	
GTF2IRD1		CFTR, CTTNBP2, EREG, LY6G6D, MUC12, QPRT, RAB27B, SLC26A3, VAV3
TCEAL1		ACSL6, AMACR, CALB1, CHN2, CTSE, CTTNBP2, GCNT3, PLCB4, PTPRO

TCF7	A1CF, ABAT, ACSL6, APCDD1, ASCL2, AXIN2, BAG2, CFTR, CHN2, CTTNBP2, DACH1, LY6G6D, MEP1A, PLK2, VAV3	ABAT, ACSL6, APCDD1, ASCL2, AXIN2, BAG2, DACH1, LY6G6D, VAV3
ASCL2	ACSL6, ANO1, APCDD1, AXIN2, CD55, CEL, CELP, CFTR, CTSE, CTTNBP2, CXCR4, DACH1, DPEP1, DUSP4, GCNT3, KIAA0226L, KRT23, LY6G6D, NOX1, PTGS2, PTPRO, QPRT, RAB27B, RARRES1, REG4, SPON1, VAV3, VNN1	CD55, CEL, CTSE, DPEP1, KIAA0226L, PTGS2, PTPRO, QPRT, REG4, SPON1, VNN1
SATB2	ACE2, ACSL6, AMACR, ANO1, ASCL2, C10orf99, CELP, CFTR, CRIP1, ENPP3, EREG, FABP1, KIAA0226L, KLK10, LRRC19, LYZ, MEIS2, NOX1, PLA2G4A, RARRES1, SDR16C5, SERPINB5, VAV3	ACE2, ACSL6, AMACR, C10orf99, CELP, CFTR, CRIP1, ENPP3, KIAA0226L, LRRC19, LYZ, MEIS2, PLA2G4A, RARRES1, VAV3
PLAGL2	A1CF, ABAT, ACE2, AGR3, ALOX5, AMACR, ANO1, ASCL2, BAG2, CALB1, CD55, CEL, CELP, CHN2, CRIP1, CTSE, DUSP4, EREG, LY6G6D, LYZ, MEP1A, PLCB4, QPRT, RAB27B, REG4, SDR16C5, SERPINB5, TACSTD2, TCN1, UGT2A3, VAV3, ZIC2	ABAT, ACE2, AGR3, AMACR, ASCL2, BAG2, CALB1, CEL, CELP, CHN2, CRIP1, DUSP4, EREG, LY6G6D, MEP1A, QPRT, SDR16C5, TCN1, UGT2A3, ZIC2

국문초록

약물 리포지셔닝과 약물 표적 동정을 위한

약물학에서의 네트워크 분석

네트워크 생물학은 생물시스템을 시스템 수준의 세포 기능을 이해하기 위해 여러 세포 내 물질들과 그들간의 관계로 이루어진 복잡계 네트워크로 표현하고 분석한다. 네트워크 개념으로 생물시스템을 살펴보는 일은 또한 약물의 세포 내 작용기전의 복잡한 양상을 드러냄으로써 신약개발 과정을 개선하는데 도움을 줄 수 있다. 네트워크 분석을 신약개발 과정에 적용하기 위해 우선 우리는 신약개발에 필요한 여러 분야의 포괄적인 지식을 모아서 약물, 타겟, 질병과 그들간의 관계들로 구성된 약물학 3-요소 네트워크 데이터베이스 (PharmDB)로 통합하였다. 두 번째로, 우리는 약물학 3-요소 네트워크 분석에 적합한 새로운 방법으로 shared neighborhood scoring 알고리즘을 개발하였다. 새로운 알고리즘을 개발한 이유는 기존의 네트워크 분석방법들은 하나의 다요소 네트워크를 여러 개의 단일요소 네트워크로 투사하여 분석함으로써 정보 손실을 유발하기 때문이다. PharmDB와 shared neighborhood scoring 알고리즘을 함께 사용하여 우리는 신규 약물 타겟 탐색, 약물의 신규 효능 및 신규 조합 탐색, 약물의

잠재적 부작용 탐색 담색을 위해 약물학 3-요소 네트워크를 분석할 수 있다. 세 번째로, 사전 지식에 기초한 네트워크 토폴로지 분석과 더불어 대용량 유전자 발현 프로파일을 네트워크 분석에 활용하는 일종의 데이터 주도 방식인 네트워크 추론을 시도하였다. 우리는 네트워크 추론 알고리즘으로 ARACNE 를 사용하여 전사 네트워크를 구축하고, 대장암에서 알려져 있는 예후예측 시그너처 유전자들을 조절하는 주요 조절자를 찾기 위해 분석하였다. 이 모든 일들은 네트워크 생물학이 더 효과적인 신약개발 과정을 위한 통합적인 해법이 될 수 있음을 보여준다.

주요어

약물학 데이터베이스, 데이터 통합, 3-요소 네트워크, 약물 타겟, 네트워크 토폴로지, shared neighborhood scoring 알고리즘, 약물 리포지셔닝, 유전자 시그너처, 대장암, 전사 네트워크, 네트워크 추론

학번: 2009-30464